Towards Clustering Hand Grasps of Individuals with Spinal Cord Injury in Egocentric Video

Mehdy Dousty, José Zariffa, *Senior Member, IEEE*

***Abstract*— Cervical spinal cord injury (cSCI) can cause paralysis and impair hand function. Existing assessments in clinical settings do not reflect an individual's performance in their daily environment. Videos from wearable cameras (egocentric video) provide a novel avenue to analyze hand function in non-clinical settings. Due to the large amounts of video data generated by this approach, automated analysis methods are necessary. We propose to employ an unsupervised learning process to produce a summary of the grasping strategies used in an egocentric video. To this end, an approach was developed consisting of hand detection, pose estimation, and clustering algorithms. The performance of the method was examined with external evaluation indicators and internal evaluation indicators for an uninjured and injured participant, respectively. The results demonstrated that a Gaussian mixture model obtained the highest accuracy in terms of the maximum match, 0.63, and the Rand index, 0.26, for the uninjured participant, and a silhouette score of 0.13 for the injured participant.**

***Clinical Relevance*— This method has the potential to allow clinicians for the first time to monitor the hand postures of individuals with cSCI at home, which could assist in remotely tailoring interventions.**

I. Introduction

A spinal cord injury (SCI) is an unexpected and devastating medical condition that changes the course of a person's life. The associated sensorimotor deficits cause individuals with SCI to have reduced ability to complete activities of daily living (ADLs), for example, eating, bathing, and dressing [1]. Moreover, people with SCI must often cope with several secondary health complications, including a neurogenic bowel and bladder, respiratory symptoms, urinary tract infections, and psychiatric issues such as reactive depression and anxiety disorders [2], [3]. Additionally, living with SCI entails considerable financial costs to individuals and the health provider. Therefore, given the 282,000 Americans and 86,000 Canadians living with SCI and the 4,300 new cases of SCI annually in Canada, effective rehabilitation assessment and treatment strategies are greatly needed [3][4].

Despite the multifaceted repercussions of cervical SCI (cSCI), studies have shown that the recovery of hand function is the first-ranked priority of the affected individuals [1]. However, current assessments of hand function, such as the Graded Redefined Assessment of Strength, Sensibility, and Prehension (GRASSP) [5], and the Toronto Rehabilitation Institute Hand Function Test (TRI-HFT) [6], are limited to a clinical or laboratory setting and are not representative of hand use in everyday contexts. In other words, the current methods do not make a distinction between capacity, which refers to an individual's highest level of function in a given task, and performance, which refers to the individual's performance of this task in their daily life [7], [8].

Wearable technology (WT) refers to devices worn directly on or attached to an individual, and has great potential for describing function in a variety of natural contexts. Accelerometers are one of the most used WTs in neurorehabilitation [9]; however, their use has focused on wrist-worn configurations that reflect arm movements, and are unable to directly gauge hand function. For example, they provide no information on the hand postures used or objects interacted with. Unlike other wearable technologies currently being investigated, wearable cameras can provide detailed data about hand use and function at home. This technology has been investigated extensively for recognition of ADLs [10], resulting in the development of techniques for automated analysis of video from such cameras. The first-person perspective (egocentric video), which captures the wearer's hand activity, can be leveraged in the monitoring and analysis of functional hand use. The feasibility of using wearable cameras in the rehabilitation context has recently been examined, with promising results [11].

While these video recordings provide valuable insight into an individual's hand functionality, manually processing the massive amounts of complex video data is prohibitively difficult, emphasizing the need for automated data analysis. One important piece of information is hand posture. Despite the considerable number of grasp taxonomies for able-bodied individuals, such as Cutkosky and Wright's [12], and Feix's [13], these classifications may not be fully applicable to SCI. The variability in grasping strategies that occurs as a result of different levels and severities of injuries complicates the classification of grasps into a defined set. One possible solution is to identify commonly used postures in an egocentric video using unsupervised learning algorithms. In this manner, an individualized taxonomy for each user could be generated, providing a summary of their grasping strategies.

In this study, a computer vision system was developed to cluster hand postures in egocentric videos.

*This study was supported in part by the Natural Sciences and Engineering Research Council of Canada (RGPIN-2014-05498), the Rick Hansen Institute (G2015-30), the Ontario Early Researcher Award (ER16-12-013), and the Craig H. Neilsen Foundation (542675).

M. Dousty and J. Zariffa are with the Institute of Biomaterials and Biomedical Engineering, University of Toronto, and KITE, Toronto Rehabilitation Institute-University Health Network. J. Zariffa is also with the Rehabilitation Sciences Institute and the Edward S. Rogers Sr. Department of Electrical and Computer Engineering, University of Toronto

II. METHODS

The sequential steps of the proposed method are hand detection, pose estimation, and clustering, as shown in Figure 1. Each of these steps is described below.

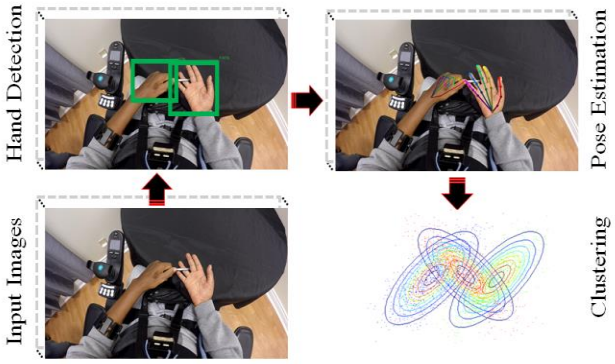

Figure 1. Steps for clustering hand postures.

A. *DATASET*

Uninjured participant: We recorded an egocentric video while a healthy participant (a 32-year-old male) was asked to grasp a ball using a spherical grasp, hold a pen using a tripod grasp, grasp a candy with a tip-to-tip pinch, and grasp a glue stick using a cylindrical grip. Figure 2 shows sample frames of each grasp. The purpose of this data was to provide a video with well-defined grasp types, which can be used as ground truth to evaluate the clustering results.

An individual with SCI: 3000 video frames were selected from an egocentric video recorded at home by an individual with cSCI (a 45-year-old female), performing several tasks using multiple hand postures. The participant had a C2 neurological level of injury, an AIS grade of A, and right and left upper extremity motor scores of 22 and 24, respectively. Figure 3 shows sample frames from this video. The purpose of this video was to provide data reflective of the intended application, with unconstrained grasping strategies.

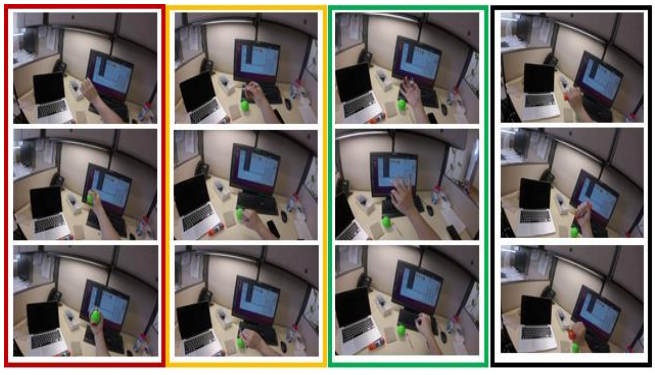

Figure 2. Sample grasps of the uninjured participant. A) Spherical grasp, B) Tripod grasp, C) Tip-to-tip pinch, D) Cylindrical grasp

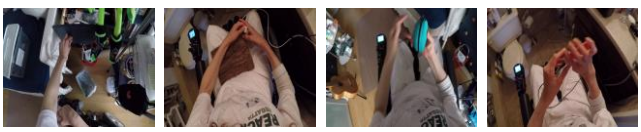

Figure 3. Sample grasps of the individual with SCI

B. HAND DETECTION AND POSE ESTIMATION

Hand detection is the first step to estimate joint coordinates. You Only Look Once (YOLO) is a regression-based object detection algorithm that enables the estimation of a large spectrum of object coordinates [14]. In this study, we used a version of YOLOv2 [15] that was previously retrained for hand detection using a dataset in which individuals with SCI interacted with various objects in a simulated home environment [16].

Next, a pose estimation algorithm can be applied to estimate hand joint coordinates. In recent years, discriminative pose estimation has made remarkable progress due to the availability of deep learning and high-quality datasets, including COCO keypoints, MPII human pose dataset, and VGG pose dataset, which has resulted in these approaches surpassing the performance of generative algorithms [17]. Here we used OpenPose [18], a discriminative approach based on convolutional pose machines [19] that demonstrates a high mean average precision for localizing hand joints. It estimates 21 joints as described in Table 1.

Table 1. 21 estimated joints in OpenPose

Joint Number	Location
1	Wrist
2	Thumb Carpometacarpal joint
3	Thumb MCP joint
4	Interphalangeal joint
5	Tip of thumb
6	First finger MCP
7	First finger Proximal interphalangeal (PIP) joint
8	First finger Distal interphalangeal (DIP) joint
9	Tip of first finger
10	Second finger MCP
11	Second finger PIP
12	Second finger DIP
13	Tip of second finger
14	Third finger MCP
15	Third finger PIP
16	Third finger DIP
17	Tip of third finger
18	Fourth finger MCP
19	Fourth finger PIP
20	Fourth finger DIP
21	Tip of fourth finger

D. Clustering and its Evaluation:

Clustering refers to the process of grouping similar entries in a data set [20]. We applied k-means [20], agglomerative hierarchical clustering algorithm (AGG) using single, ward, and complete linkage with Euclidean distance [21], Gaussian mixture model (GMM) [22], clustering using representatives (CURE) [23], balanced iterative reducing and clustering using hierarchies (BIRCH) [24], and density-based spatial clustering of applications with noise (DBSCAN) [25].

One of the fundamental elements for implementing any clustering method is the number of classes. Here, we used the elbow method to extract the number of clusters based on the Bayesian information criterion (BIC) [26].

Furthermore, to test the validity of the clustering algorithms, one can use internal evaluation indicators (IEI) and external evaluation indicators (EEI). IEI examines the relationship between the data points within and between clusters. This process does not require any data labeling. In

EEI, clustering performance is evaluated by making a comparison between the clustering results and the true class labels [20].

Maximum match (MM), Jaccard coefficient (JC), Rand index (RAND), and Fowlkes-Mallows (FLK) index were used to quantify EEI. MM quantifies the extent to which each cluster contains points only from one partition. It varies between 0-1, where 1 indicates perfect clustering. JC measures the performance of a cluster based on the intersection over the union of the true and estimated clusters. RAND is a similarity measurement between estimated clusters and labeled clusters computed by considering all pairs of instances between two clusters, where 1 indicates perfect clustering. Finally, FLK quantifies the geometric average between recall and precision. A higher number indicates a greater similarity between partitions and clusters. On the other hand, the silhouette score was used to quantify IEI. It measures the similarity of a data point to others in its cluster compared to other clusters. The silhouette has a range between -1 to +1, with high values indicating good clustering while low values show poor clustering [20].

III. Results

A. Uninjured participant

In 45% of the frames in the uninjured participant dataset, the OpenPose algorithm was unable to identify all the joint locations with a confidence score greater than 0.2. To mitigate this difficulty, we constrained the joints included in the clustering analysis to joints 1-15 in Table 1, for two reasons. First, these joints tend to be less occluded during object interactions. Second, the four hand postures described above primarily involve the thumb, index, and middle finger. We therefore selected the frames in which the finger joints 1 to 15, described in Table 1, had at least 0.2 coordinate confidence. This threshold was selected empirically. Next, we clustered the frames into four clusters. Table 2 shows each clustering method's performance in the uninjured participant.

Table 2. Clustering performance based on EEI for the uninjured participant.

Clustering /EEI	*MM*	*JC*	*Rand*	*FLK*
K-means	0.48	0.08	0.12	0.40
Agg Ward L2	0.54	0.06	0.20	0.42
Agg Single L2	0.31	0.15	0.00	0.48
Agg Complete L2	0.39	**0.16**	0.05	0.43
BIRCH	0.54	0.06	0.20	0.42
CURE	0.30	0.14	0.00	**0.50**
DBSCAN	0.46	0.09	0.10	0.41
GMM	**0.63**	0.10	**0.26**	0.46

The results for the uninjured participant reveal that among all of the applied clustering algorithms, GMM outperforms the other methods in most of the EEI. The confusion matrix for the GMM is shown in Table 3, where C is the predicted label and T is the true label.

Table 3. Confusion matrix for GMM clustering in the uninjured participant dataset.

	C1	**C2**	**C3**	**C4**
T1	**139**	51	18	6
T2	51	**240**	14	32
T3	86	14	**170**	3
T4	56	76	18	**165**

One essential element of a clustering algorithm is the number of clusters. In the results above, we assumed that the number of clusters is known; however, in practice, there is no intuition about them. Here, we used the elbow method to extract the number of clusters. Figure 4 shows the elbow method based on the Bayesian information criterion (BIC) for the GMM. It is clear from the figure that after five clusters the gradient of BIC plateaus, which suggests 4 clusters exist in the data.

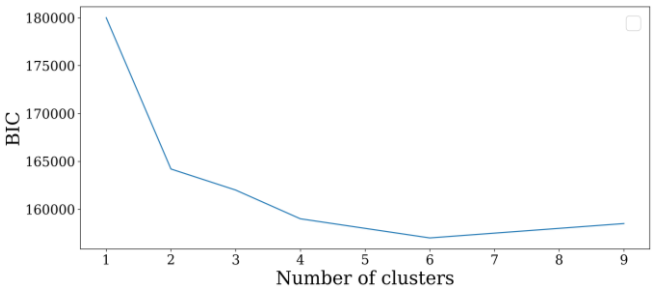

Figure 4. BIC for the different number of clusters in the uninjured participant.

B. Injured subject

Using the first 15 joints coordinate, we computed the optimum number of clusters based on BIC of the GMM clustering algorithm, depicted in Figure 5. We used 100 random initializations for the GMM to avoid missing a globally optimal solution.

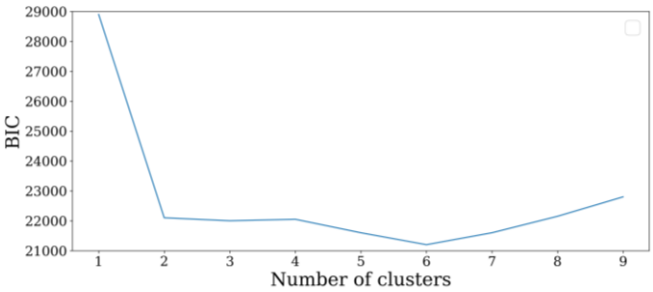

Figure 5. BIC for various number of clusters in an individual with SCI

Figure 5 suggests that the optimal number of clusters is 5 or 6, meaning that the participant was using 5 or 6 distinct hand postures in the video. This results aligns well with direct observation of the data, through which we determined that the subject demonstrated 6 main hand postures. We examined the clustering performance using EEI with the silhouette score and acquired 0.14, 0.13 for 5 and 6 clusters respectively. In summary, the results based on the uninured participant and an individual with SCI show the potential of the proposed method to cluster hand posture.

IV. Discussion

Monitoring the hand function of individuals with SCI in their daily living environments will lead to a more holistic assessment of the impact of new interventions, as well as allow clinicians to provide more effective remote care. Here, we proposed to use a wearable camera in conjunction with computer vision algorithms to distill complex egocentric videos into summaries of grasping strategies. These initial

results are the first proof-of-concept demonstration of clustering hand grasps based on pose estimation data.

Despite the potential of the method, the accuracy of this novel paradigm needs improvement before it can be deployed in real-world applications. First, the current pose estimation algorithms are unable to estimate hand joints precisely in heavily occluded scenarios. This situation most often happens when an individual is interacting with objects, such that portions of the hand are not visible. One possible remedy to this problem may be to analyze hand configurations towards the end of the reaching stage when the hand is being positioned into the desired grasp type but before it is hidden by the object. Another possibility would be to merge a joint tracking algorithm with the current pose estimation method, in order to help predict joint locations when the pose estimation generates low confidence scores. We postulate that by acquiring more accurate pose estimation, the clustering performance would increase. Second, we only applied the clustering on spatial information at the level of each frame, and also annotated hand postures at the frame level. However, any hand object interaction comprises a sequence of movements, so defining the number of postures during an object interaction can be challenging. One possible solution that may increase the clustering performance would be to use spatio-temporal clustering methods to cluster hand trajectories. Action respecting embedding [27] and spatio-temporal graph-based manifold embedding [28] are two example methods that allow for clustering repetitive actions by analyzing embedding manifolds. Finally, the current pose estimation algorithm is unable to estimate the wrist angle, which is an important piece of postural information after cSCI. Including wrist angle information may further improve the clustering.

This is the first attempt to summarize the hand postures used after cSCI by using wearable technology in non-clinical environments. These pilot results demonstrate the successful clustering of similar hand postures. In the future, we will increase the accuracy of the proposed paradigm by addressing the discussed limitations.

V. References

[1] K. D. Anderson, "Targeting Recovery: Priorities of the Spinal Cord-Injured Population," *J. Neurotrauma*, 2004.

[2] S. W. Lim *et al.*, "Anxiety and Depression in Patients with Traumatic Spinal Cord Injury: A nationwide population-based cohort study," *PLoS One*, 2017.

[3] J. S. Krause and L. L. Saunders, "Health, secondary conditions, and life expectancy after spinal cord injury," *Arch. Phys. Med. Rehabil.*, 2011.

[4] A. Farry and D. Baxter, "The incidence and prevalence of spinal cord injury in Canada: overview and estimates based on current evidence," *Rick Hansen Inst. Urban Futur. Inst.*, pp. 1–49, 2010.

[5] S. Kalsi-Ryan *et al.*, "The Graded Redefined Assessment of Strength Sensibility and Prehension: Reliability and Validity," *J. Neurotrauma*, 2012.

[6] N. Kapadia, V. Zivanovic, M. Verrier, and M. Popovic, "Toronto Rehabilitation Institute–Hand Function Test: Assessment of Gross Motor Function in Individuals With Spinal Cord Injury," *Top. Spinal Cord Inj. Rehabil.*, 2012.

[7] R. J. M. Lemmens, A. A. A. Timmermans, Y. J. M. Janssen-Potten, R. J. E. M. Smeets, and H. A. M. Seelen, "Valid and reliable instruments for arm-hand assessment at ICF activity level in persons with hemiplegia: A systematic review," *BMC Neurol.*, 2012.

[8] R. J. Marino, "Domains of outcomes in spinal cord injury for clinical trials to improve neurological function.," *J. Rehabil. Res. Dev.*, vol. 44, no. 1, pp. 113–122, 2007.

[9] K. J. Waddell *et al.*, "Does Task-Specific Training Improve Upper Limb Performance in Daily Life Poststroke?," *Neurorehabil. Neural Repair*, 2017.

[10] T. H. C. Nguyen, J. C. Nebel, and F. Florez-Revuelta, "Recognition of Activities of Daily Living with Egocentric Vision : A Review," *Sensors (Switzerland)*, no. December 2015, 2016.

[11] J. Likitlersuang, E. R. Sumitro, T. Cao, R. J. Visée, S. Kalsi-Ryan, and J. Zariffa, "Egocentric video: A new tool for capturing hand use of individuals with spinal cord injury at home," *J. Neuroeng. Rehabil.*, 2019.

[12] M. R. Cutkosky, "On Grasp Choice, Grasp Models, and the Design of Hands for Manufacturing Tasks," *IEEE Trans. Robot. Autom.*, 1989.

[13] D. Kragic, A. M. Dollar, J. Romero, T. Feix, and H.-B. Schmiedmayer, "The GRASP Taxonomy of Human Grasp Types," *IEEE Trans. Human-Machine Syst.*, vol. 46, no. 1, pp. 66–77, 2015.

[14] J. S. D. R. G. A. F. Redmon, "(YOLO) You Only Look Once," *Cvpr*, 2016.

[15] J. Redmon and A. Farhadi, "YOLO9000: Better, faster, stronger," in *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, 2017.

[16] R. J. Visée, J. Likitlersuang, and J. Zariffa, "An Effective and Efficient Method for Detecting Hands in Egocentric Videos for Rehabilitation Applications," *arXiv Prepr.*, Aug. 2019.

[17] Z. Cao, T. Simon, S. E. Wei, and Y. Sheikh, "Realtime multi-person 2D pose estimation using part affinity fields," in *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, 2017.

[18] T. Simon, H. Joo, I. Matthews, and Y. Sheikh, "Hand keypoint detection in single images using multiview bootstrapping," *Proc. - 30th IEEE Conf. Comput. Vis. Pattern Recognition, CVPR 2017*, vol. 2017-Janua, pp. 4645–4653, 2017.

[19] S. E. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh, "Convolutional pose machines," *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 2016-Decem, pp. 4724–4732, 2016.

[20] D. Xu and Y. Tian, "A Comprehensive Survey of Clustering Algorithms," *Ann. Data Sci.*, vol. 2, no. 2, pp. 165–193, 2015.

[21] F. Murtagh and P. Contreras, "Algorithms for hierarchical clustering: An overview," *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.*, 2012.

[22] J. C. Spall and J. L. Maryak, "A Feasible Bayesian Estimator of Quantiles for Projectile Accuracy From Non-iid Data," *J. Am. Stat. Assoc.*, 1992.

[23] S. Guha, R. Rastogi, and K. Shim, "CURE: An efficient clustering algorithm for large databases," *Inf. Syst.*, 2001.

[24] T. Zhang, R. Ramakrishnan, and M. Livny, "BIRCH: An Efficient Data Clustering Method for Very Large Databases," *SIGMOD Rec. (ACM Spec. Interes. Gr. Manag. Data)*, 1996.

[25] M. Ester, H.-P. Kriegel, J. Sander, X. Xu, and X. Ester, M., Kriegel, H. P., Sander, J., & Xu, *A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise*. 1996.

[26] D. Pelleg and A. Moore, "X-means: Extending K-means with Efficient Estimation of the Number of Clusters," *CEUR Workshop Proc.*, 2015.

[27] M. Bowling, A. Ghodsi, and D. Wilkinson, "Action respecting embedding," in *ICML 2005 - Proceedings of the 22nd International Conference on Machine Learning*, 2005.

[28] M. Al Ghamdi and Y. Gotoh, "Alignment of nearly-repetitive contents in a video stream with manifold embedding," *ICASSP, IEEE Int. Conf. Acoust. Speech Signal Process. - Proc.*, pp. 1255–1259, 2014.