POLITECNICO
MILANO 1863

SCUOLA DI INGEGNERIA INDUSTRIALE
E DELL'INFORMAZIONE

Executive Summary of the Thesis

A multi-modal approach combining first person vision and surface EMG for the functional assessment of the upper extremities: a feasibility study

Laurea Magistrale in Biomedical Engineering - Ingegneria Biomedica

Author: Giada Zecchin

Advisor: Prof. Emilia Ambrosini

Co-advisor: Andrea Bandini

Academic year: 2022-2023

1. Introduction

Upper extremity (UE) function recovery is essential to improve the quality of life and it is a top priority for people with spinal cord injury (SCI) and stroke. Rehabilitation is crucial to promote brain plasticity and regain lost function, however, improvement profiles might change over time and do not always follow a predictable pattern.

Various assessment scales, such the Fugl-Meyer Assessment (FMA) and the Action Research Arm Test (ARAT) are commonly used to evaluate UE function in stroke patients. Scales used to assess UE function in SCI patients include the International Standards for Neurological Classification of Spinal Cord Injury (ISNCSCI), the Spinal Cord Independence Measure (SCIM), and Graded Redefined Assessment of Strength, Sensibility and Prehension (GRASSP). Although such scales are useful for measuring functional capacity, they may not accurately assess real-world performance due to their subjective nature and lack of specificity.

Recovering the use of UE can be challenging for several reasons: financial pressure on the healthcare system which results in early discharge, lack of outcome measures that take into account the performance (and not only capacity) mainly at home, and long distances.

Recent years have seen advancements in daily life assessment and rehabilitation using wearable technologies like accelerometers and inertial measurement units (IMU). However, these methods fail to capture complex hand and finger movements [1]. Additionally, patients with spasticity or a limited range of motion may find it difficult to use such devices.

First-person vision (FPV) technology combined with deep learning algorithms has been developed for detecting the hands, objects, and interactions and to monitor hand movements in real-life contexts [2]. Through FPV it is possible to capture the user's point of view and, if the camera is worn on the head, to focus on hand movements and manipulation. Compared to third-person vision, where the camera is fixed and "watches" the person from an external perspective, FPV offers a more realistic and engaging experience. Previous studies have investigated the feasibility and validity of using FPV technology to measure hand use and function in individuals with SCI and in stroke survivors [2][3]. However, the main drawback is that, although accurate in detecting hand-object interactions, FPV fails to discriminate interactions in presence of "non-standard" grasps and compensatory strategies. The research hypothesis is that such limitation can be overcome by combining FPV with surface electromyography (sEMG). sEMG is a non-invasive technique that involves placing small electrodes on the surface of the skin to detect electrical signals from muscle contractions and it has been used in various studies to investigate muscle activation patterns and to track progress in therapy [4].

Therefore, this thesis aims to develop a multimodal classification algorithm that combines FPV and sEMG to identify functional hand-object interactions during daily life activities. To evaluate the feasibility of the multi-modal approach, data were collected on healthy subjects during standardized tasks resembling activities of daily living (ADLs) in a lab setting.

2. Materials and Methods

2.1. Participant and Experimental Protocol

The study was conducted on 10 healthy adults (6 males and 4 females with an average age of 27) at the Biorobotics Institute, Scuola Superiore Sant'Anna, Pisa (Italy) and was approved by the Joint Ethical Committee of Scuola Superiore Sant'Anna and Scuola Superiore Normale (Nr. 3/2023). Participants wore a sleeve with dry sEMG sensors, and a wearable camera mounted on their foreheads. We used the following equipment: OTBioelettronica Sessantaquattro device (OTBioelettronica, Turin, Italy) for sEMG acquisition (sampled at 2k Hz), an array of 32 dry MXene electrodes [5] and a GoPro Hero 8 (GoPro Inc., California), that recorded footage at 30 frames per second with a resolution of 1920x1080 pixels. The sEMG amplifier included an auxiliary channel (AUX), connected to a trigger circuit including a button and two LEDs. The button was pressed at the start and end of each task. As a result, power was supplied to the LEDs, and 3.3mV-pulses were transmitted to the AUX. This LED-activated switch provided offline synchronization between sEMG data and videos.

The experimental protocol involved the simultaneous recording the of sEMG signal and videos of activities. Each session, lasting 60 minutes, consisted of three phases, including setting up the sEMG system on the subject's forearm, setting up the camera, and recording task execution.

In this study, following the strategy from Bandini et al. (2022) [2], we defined functional interaction as a manipulation of the object, any contact with a fixed or portable object, while non-functional interaction referred to no contact between the hand and the object, any self-contact and any contact with another person. Each subject was asked to perform the following set of functional tasks (that were repeated three times, with each repetition lasting for five seconds followed by a five-second resting period): box and block test, pouring water, writing a sentence, and typing a message on a smartphone. The maximum voluntary contraction (MVC) was also performed and the resulting data were used for sEMG signal normalization (MVC data were not used for the detection of functional / non-functional interactions).

2.2. Data Analysis

Using Matlab R2022b, we conducted data analysis according to the following pipeline:

1. Manual labelling
2. sEMG and FPV synchronization
3. Dataset split
4. Interaction detection from sEMG
   (a) Feature extraction
   (b) Feature selection
   (c) sEMG performance
5. Interaction detection from FPV
   (a) Detection extraction
   (b) Side correction
   (c) FPV performance
6. Interaction detection by combining sEMG and FPV
   (a) Performance comparison (single-modal vs multi-modal)

2.2.1 Manual labelling

Manual labelling was performed from each video to identify:

- the ground truth of the hand-object interaction state (i.e. functional interaction (=1) or non-functional interaction (=0));

- the ground truth of the bounding box coordinates of the hands, and the detected hand side (left or right);
- the beginning and end frames of each task (i.e., frames with the LED switched on).

2.2.2 sEMG and FPV synchronization

A script was created to automatically align the spike signal from the trigger channel with the corresponding frame that detected the LED. This enabled the synchronization of ground truth frames with the sEMG samples.

2.2.3 Dataset split

The performance of the hand-object interaction detection was evaluated on sEMG, FPV, and their combinations. 70% of the data (i.e., 7 randomly selected participants) was used for training and validating the algorithms, while the remaining 30% (3 participants) was used as the test set. The dataset split was repeated 12 times to assess the approach's robustness to variations in the data across subjects.

2.2.4 Interaction detection from sEMG

Feature extraction First, inter-task periods (i.e., intervals when the participants prepared themselves for the next task) were removed.
A Notch filter (2nd-order infinite impulse response (IIR) bandstop filter) with a cut-off frequency of 49 Hz and 51 Hz was applied to remove the power line interference of 50 Hz. The signal was then filtered with a Butterworth bandpass filter at 10-250Hz.
Each channel was normalized to the maximum peak achieved during the MVC recordings. The highest amplitude value was obtained by rectifying the signal and calculating the maximum value over a 100 ms moving average window.
Principal Component Analysis (PCA) was then applied, projecting the 32 signals onto the first 11 principal components, which accounted for over 90% of the total variance of the data. This choice of 11 components was justifiable given that there are approximately 11 detectable superficial forearm muscles.
The sEMG signals were windowed using four different window lengths (250ms, 500ms, 750ms, and 1s) with 50% overlap. From each window, twelve time-domain (TD) and nine frequency-domain (FD) features were extracted. Six time-domain features were generated based on the rectified signal amplitude, while the other six required the signal to be enveloped (using a moving average with a window of 100 ms, i.e. 200 samples) before feature extraction. A Fast Fourier transform was used to convert the sEMG signal in the frequency domain and extract FD features from the resulting power spectral density (PSD). Some of the features included integrated EMG (IEMG), average amplitude change (AAC), mean absolute value (MAV), median absolute deviation (MAD), waveform length (WL), log detector (LD), root mean square (RMS), variance (VAR), simple square integral (SSI), coefficient of variation (COV), kurtosis (KURT), skewness (SKEW), mean frequency (MNF), median frequency (MDF), total power (TP), mean power (MP), peak frequency (PKF), first, second, and third spectral moments (SM1, SM2, SM3), and variance of central frequency (VCF). Since 21 features were extracted from 11 PCs, a total of 231 features were achieved for each time window.
The ground truth of the interaction state, represented by an array of 1s and 0s (see section 2.2.1), was also divided into windows, and a value of 1 was assigned if the majority of samples indicated functional interaction and 0 otherwise.

Feature selection A feature ranking was performed on the training set using the Maximum Relevance Minimum Redundancy (MRMR) algorithm, which identified the most important and non-redundant features. We selected the top 6, 12, and 23 most predictive features, representing 2.5%, 5%, and 10% of the total number of features respectively.

sEMG performance Four classifiers were compared for detecting the hand-object interaction state from sEMG features: support vector machines (SVM) with linear and radial basis function (RBF) kernels, random forest (RF), and k-nearest neighbour (kNN). The selected features were the predictors, and the ground truth of the interaction state served as the response variable. Hyperparameter tuning involved testing various combinations of hyperparameters and selecting the one that yielded the highest accuracy and F1-score. The Leave-One-Subject-Out Cross-Validation (LOSO-CV) method was used for validation, and the best-performing hyperparameters were applied to the test set.

Training and testing were performed only on one combination of the dataset split to identify the best algorithm, which was then trained on the other 12 combinations. The sEMG classification performance was evaluated using accuracy, F1-score, precision, recall, and specificity metrics.

2.2.5 Interaction detection from FPV

Detection extraction The video clips were split into frames and Shan et al.'s deep learning model [6], which implements a Faster-RCNN, was used to locate the hands, determine their contact state, and identify the object they are in contact with. The detection results were saved in CSV files, one file per video, with 12 columns containing information about detected hands and objects, their bounding box coordinates, confidence scores, contact states (0:no contact, 1:self contact, 2:contact with another person, 3:contact with portable object, 4:contact with fixed object), offset vectors (offset between the hand and the object), and hand side (left or right).

Just like the sEMG signal processing (see paragraph 2.2.4), the inter-task periods were removed using the manually marked start-end frames of each task.

Due to the presence of at least two hands (we ensured not to interfere with the subject's task execution so that only his right and left hands were within the camera view) and objects, each frame resulted in more than one detection. Therefore, a series of processing steps were required to obtain a maximum of two detections per frame (i.e., one for each hand). The first step involved removing the object detections to focus on the hand detections. Then the hand detections with a confidence value lower than 0.5 were discarded.

Side correction To remove duplicated hand sides (i.e., detection in which left and right hands were not differentiated), a side correction step was implemented, using four different criteria. The first three criteria were: (1) considering only the duplicated detection with the highest confidence score; (2) assigning "right" ("left") to the right-most (left-most) detection; and (3) retaining the duplicated detection with the uppermost coordinates. A fourth criterion was based on machine learning algorithms that were trained to recognize the correct side based on the coordinates of the hand bounding boxes.

FPV performance We selected as functional interactions (label=1) the frames where the contact state was either 3 or 4 (i.e., contact with a movable and fixed object, respectively), whereas the frames with a contact state < 3 belonged to the non-functional interaction class (label=0) (see section 2.1 and [2]).

Finally, the interaction state output and the relative manual ground truth were processed similarly to the sEMG signal (see section 2.2.4) using consecutive windows of different duration (250 ms, 500 ms, 750 ms, and 1 s) with a 50% overlap. A value of 1 was assigned if the majority of frames indicated functional interaction and 0 otherwise. By comparing the interaction state of the test set with the manually identified ground truth, we were able to obtain the performance of the FPV in detecting hand-object interactions, also evaluated in terms of accuracy, F1-score, precision, recall, and specificity metrics.

2.2.6 Interaction detection by combining sEMG and FPV

We then combined the sEMG and FPV modalities using three different approaches. The first (FPV + EMG) involved concatenating the interaction state from the FPV analysis with the most relevant features extracted from the sEMG signal. We proceed to train the machine learning algorithm that performed best in the sEMG single-modal approach. The second (FPV AND EMG) and third (FPV OR EMG) approaches used the AND and OR logical operators, respectively, to combine the interaction state predicted by the sEMG signal and FPV analysis. The classification performance combinations were evaluated by comparing the predicted interaction state with the relative ground truth, using the same five metrics as those used for the single-modal approaches.

Performance comparison (single mode vs multimodal) The non-parametric Kruskal-Wallis test was used to compare the classification performance of the different approaches (single and multi-modal). If the obtained p-value was lower than 0.05, a post-hoc analysis was performed using Dunn's test with the Bonferroni correction. Two statistical analyses were conducted: one compared the three algorithms for combining sEMG and FPV, and the other compared the best-performing combination method with two single-modal approaches.

3. Results

3.1. sEMG Results

PCA In order to reduce data dimensionality we considered only the first 11 PCs as, on average, they covered over 92.2% of the variance.

sEMG feature selection The top six features ranked by the MRMR algorithm were: WL, MDF, KURT, MNF, SM3 and PFK.

sEMG performance Table 1 depicts the classifier and the relative training and validation accuracy and F1-score. The SVM model trained with the RBF kernel, using the top six sEMG features, and the 1-second window length had the best performance.

N	Classifier	Accuracy	F1-score
6	**SVM (RBF)**	**0.685**	**0.720**
6	SVM (linear)	0.616	0.587
6	RF	0.679	0.703
6	kNN	0.650	0.669
12	SVM (RBF)	0.666	0.683
12	SVM (linear)	0.601	0.584
12	RF	0.661	0.683
12	kNN	0.618	0.640
23	SVM (RBF)	0.667	0.681
23	SVM (linear)	0.608	0.590
23	RF	0.676	0.696
23	kNN	0.625	0.650

Table 1: Number of features (N), classifier, accuracy and F1-score of the sEMG training and validation performance, tested on window length of 1 second. The best classification performances are highlighted in bold.

As for the testing on the 1-second window length test set, the sEMG approach achieved an accuracy of 0.660, an F1-score of 0.708, a precision of 0.626, a recall of 0.814 and a specificity of 0.501.

3.2. FPV Results

The videos in the study, for each subject, were about 8.5 minutes long on average and had a total of about 15,300 frames, which were reduced in resolution to 848x480 pixels.

Side correction The second criterion (i.e., assigning "right" ("left") to the right-most (left-most) detection) had the highest scores, with an accuracy value of 0.91. It was also chosen because it is the most suitable for daily life activity recording.

FPV performance Based on the F1-score, the best results were obtained with the 1-second window length. The five metrics values that described FPV performance in detecting hand-object interactions were the following: the accuracy value was 0.528, the F1-score was 0.682, the precision was 0.518, the recall was 0.997, and the specificity was 0.047. In particular, FPV achieved high recall values but had very low specificity.

3.3. Combination Performance

Comparison of combination techniques Table 2 reports the results of the three different methods implemented for combining FPV and sEMG predictions. Significant differences were found in accuracy, precision, and specificity among the three multimodal approaches, while none were observed in the F1-score. The "FPV OR EMG" combination performed significantly worse than both "FPV + EMG" and "FPV AND EMG" which, on the other hand, had similar results. However, in terms of recall, "FPV OR EMG" significantly outperformed both "FPV + EMG" and "FPV AND EMG".

Single-modal vs multimodal approach The "FPV + EMG" method showed the lowest interquartile range for both accuracy and F1-score among the combinations, indicating less variability in the data. Therefore it was selected for the following analysis which compared the two single modalities to the "FPV + EMG" (Table 3).

Metric	Median (interquartile range)			Kruskal-Wallis p-value	Post-hoc p-value		
	FPV+EMG	FPV AND EMG	FPV OR EMG		A	B	C
Accuracy	0.716 (0.067)	0.719 (0.080)	0.598 (0.043)	<0.01	1	<0.01	<0.01
F1score	0.743 (0.061)	0.723 (0.092)	0.717 (0.030)	0.055	/	/	/
Precision	0.712 (0.075)	0.722 (0.059)	0.560 (0.036)	<0.01	1	0.01	<0.01
Recall	0.849 (0.118)	0.771 (0.124)	0.998 (0.004)	<0.01	0.525	0.01	<0.01
Specificity	0.631 (0.143)	0.665 (0.133)	0.165 (0.069)	<0.01	0.969	<0.01	<0.01

Table 2: Statistical analysis on three combination options. A= FPV + EMG compared to FPV AND EMG; B= FPV + EMG compared to B=FPV OR EMG; C= FPV AND EMG compared to FPV OR EMG. Significant difference p-value=0.05

Metric	Median (interquartile range)			Kruskal-Wallis p-value	Post-hoc p-value		
	FPV	EMG	FPV+EMG		D	E	F
Accuracy	0.650 (0.067)	0.653 (0.044)	0.716 (0.067)	0.003	1	0.005	0.016
F1score	0.745 (0.037)	0.698 (0.068)	0.743 (0.061)	0.017	0.082	1	0.023
Precision	0.598 (0.048)	0.635(0.036)	0.712 (0.075)	<0.01	0.114	<0.01	0.173
Recall	0.990 (0.007)	0.776 (0.125)	0.849 (0.118)	<0.01	<0.01	<0.01	1
Specificity	0.297 (0.146)	0.520 (0.081)	0.631 (0.143)	<0.01	0.002	<0.01	0.913

Table 3: Statistical analysis single-modal vs multi-modal approach. D=FPV compare to EMG; E=FPV compare to the multimodal approach FPV+EMG; F=EMG compare to the multimodal approach FPV+EMG. Significant difference p-value=0.05

Based on the Kruskal-Wallis test, we saw that there were significant differences between the three approaches for all metrics. The multi-modal approach had higher scores for accuracy, precision, and specificity, while FPV alone performed better in terms of recall. The multimodal system had a significantly higher F1-score than the single EMG and a similar one compared to the FPV approach.

4. Discussion

We introduced and validated, on healthy adults, a novel multi-modal approach that integrates FPV and sEMG to automatically detect hand-object interactions.

As for the sEMG classification performance, the SVM with RBF kernel (which can capture the non-linear correlations between the data points with more intricate decision limits), the 1-second window length, and the first six selected features (WL, MDF, KURT, MNF, SM3, PFK) best predicted functional interactions. Such features can capture various aspects of the sEMG signal related to muscle like activation, endurance, contraction, and asymmetry during hand-object interactions. As suggested by the high value of most of the metrics, the sEMG approach was able to accurately distinguish between functional and non-functional interactions to a certain extent. However, the system had low specificity, which indicates its lower ability to identify non-functional interactions.

The analysis of FPV classification performance revealed that the 1-second time window length was the most effective, similar to EMG. The high recall suggested that the FPV performed well in recognizing functional interactions. However, it struggled to identify non-functional interactions, as demonstrated by the very low specificity. We were interested in detecting hand-object interactions. These occur over a span of a few seconds rather than milliseconds, which may explain why the longer window length of 1 second performed better in both single-modal approaches.

From the non-parametric Kruskal-Wallis test and post-hoc analysis, we observed that the "FPV + EMG" and "FPV AND EMG" combinations had comparable performance. Even though "FPV OR EMG" was successful in identifying almost all hand-object interactions, as demonstrated by the very high recall, it had the poorest performance across every other metric. Both the "FPV + EMG" and "FPV AND EMG" combinations yielded encouraging results in terms of correctly predicting most functional interactions, as indicated by the high values across most metrics. In addition, the specificity of the multi-modal analysis improved compared to both single-modal approaches. These findings suggest that the combination of FPV technology and sEMG analysis is an effective approach for capturing functional and non-functional interactions.

Previous studies ([2][3]) have investigated the feasibility and validity of using egocentric video technology to measure hand use and function in individuals with SCI and stroke survivors living in the community. The studies found that the technology accurately captured hand movements and hand-object interactions, suggesting that it is an effective method for analyzing hand use and hand roles during daily activities in these populations. However, FPV is unable to differentiate between different types of interactions when unconventional grasping techniques and compensatory methods are used, therefore a multimodal approach can be helpful in addressing such issue.

Limitations and Future Work The study had several limitations, including a small sample size, limited testing on healthy individuals, and a focus on specific tasks that may not capture the full range of hand movements. Additionally, issues such as prolonged acquisition time, short battery life, and extensive manual labeling were encountered. Some proposed solutions for future work include expanding the sample size, testing the approach on individuals with hand impairments, evaluating performance outside the clinical setting, optimizing the wearable sleeve design, and distributing the labeling task to more than one researcher.

5. Conclusion

The study found that using a combination of FPV and sEMG is an effective way to capture hand-object interactions in healthy individuals. This information is important because understanding how the hand interacts with objects can help tailor therapy and maximize treatment outcomes. The sEMG single-modal approach shows promising results for both types of interactions but needs improvement in reducing false positives and negatives. The FPV single-modal approach can identify functional interactions but is not as good at detecting the non-functional ones, as shown by low specificity. The choice of combination significantly impacts performance and, overall, the multi-modal approach resulted in positive outcomes for all five evaluation parameters. Future research will involve validating our findings on subjects with SCI or stroke, both in a clinical setting and in their homes.

6. Bibliography

References

[1] Peter S. Lum, Liqi Shu, Elaine M. Bochniewicz, Tan Tran, Lin Ching Chang, Jessica Barth, and Alexander W. Dromerick. Improving Accelerometry-Based Measurement of Functional Use of the Upper Extremity After Stroke: Machine Learning Versus Counts Threshold Method. *Neurorehabilitation and Neural Repair*, 34(12):1078–1087, 12 2020.

[2] Andrea Bandini, Mehdy Dousty, Sander L. Hitzig, B. Catharine Craven, Sukhvinder Kalsi-Ryan, and José Zariffa. Measuring Hand Use in the Home after Cervical Spinal Cord Injury Using Egocentric Video. *Journal of neurotrauma*, 39(23-24):1697–1707, 12 2022.

[3] Meng Fen Tsai, Rosalie H. Wang, and Jose Zariffa. Identifying Hand Use and Hand Roles after Stroke Using Egocentric Video. *IEEE Journal of Translational Engineering in Health and Medicine*, 9, 2021.

[4] Negin Hesam-Shariati, Terry Trinh, Angelica G. Thompson-Butel, Christine T. Shiner, and Penelope A. McNulty. A Longitudinal Electromyography Study of Complex Movements in Poststroke Therapy. 1: Heterogeneous Changes Despite Consistent Improvements in Clinical Assessments. *Frontiers in Neurology*, 8, 7 2017.

[5] Nicolette Driscoll, Brian Erickson, Bren-

dan B Murphy, and Andrew G Richardson. MXene-infused bioelectronic interfaces for multiscale electrophysiology and stimulation. 2021.

[6] Dandan Shan, Jiaqi Geng, Michelle Shu, and David F Fouhey. Understanding Human Hands in Contact at Internet Scale. Technical report, 2020.