Stroke Rehabilitation Therapy for the Upper Extremity with a Virtual Coach for Compensation Reduction

Ana Rita Mendes Cóias
ana.coias@tecnico.ulisboa.pt

Instituto Superior Técnico, Lisboa, Portugal

February 2021

Abstract

Stroke is one of the leading causes of death and long-term disability in the western world. The increasing demands concerning stroke rehabilitation and in-home exercise promotion increased the need for adequate, affordable, and accessible Assistive Technology (AT) to promote patients' compliance in therapy while exercising autonomously. Independent exercise with an assistive system requires objective methods to assess patients' quality of movement and track their progress. In this work, we develop an image-based Virtual Coach (VC) capable of monitoring upper extremity rehabilitation exercises focused on motor compensation reduction. The VC proposes three exercises and assesses users' compensation patterns from images acquired with a webcam. It provides proper visual and audio feedback and instructions through a User Interface (UI). We propose objective measures and classification approaches - a Rule-based (RB) and a Neural Network (NN) based - to assess motor compensation patterns from 2D positional data for the three exercises. For exercise 1, the RB approach assessed different compensation patterns with F_1 score of 76.69%. For exercise 2 and 3, the NN based approach revealed F_1 score of 72.56% and 79.87%, respectively. A group of seven volunteers exercised with the VC in a small experimental session. The group found the system enjoyable and relevant for rehabilitation administration. These results give evidence about the value of this kind of system to aid stroke patients under rehabilitation and accurate performance assessment from 2D data. This latter enables to automate rehabilitation programs monitorization in any device with a 2D camera, such as tablets, smartphones, or robotic assistants.
Keywords: Stroke, Virtual Coach, Performance Assessment, 2D Positional Data

1. Introduction

With the escalating demands towards stroke rehabilitation and the increase of in-home exercise recommendations [5], the need for new means to evaluate patients' motor performance has risen [12, 15]. In conventional assessment tests, therapists assess movement quality based on observation, thus being highly subjective [12]; with the degree of experience implying distinct treatment approaches [15]. Quantitative and objective methods allow patients' progress tracking, impaired movements' understanding, and formulation of standard therapy regimens [12, 14]. Also, in this scenario, assistive systems, such as Virtual Coaches, can aid patients' exercise autonomously. These systems are required to be adequate, affordable, and accessible, with a complex interaction model to keep the user engaged [19, 2, 6],

Patients' physically impaired often exhibit compensation behaviors to accomplish a task. Motor compensation is the presence of new movement patterns derived from the adaptation or substitution of old ones, which might help patients' execute a task [20, 13]. New patterns can include the use and activation of additional or new body joints and muscles. Most typical compensation behaviors are trunk displacements, rotation, and shoulder elevation. These functional strategies are commonly observed in reaching and are highly related to severe impairment levels [13].

Early in the recovery process, the use of compensation strategies promotes patients' upper limb participation in task performance. However, their persistence may obstruct real motor function recovery and must be reduced during therapy through appropriate exercise instructions [13].

In this work, we present a method to assess quantitatively motor compensation from video frames during upper limb exercise performance. We created a labelset (Table 4) for each video frame of the dataset regarding the observed compensation patterns. We then explore two methods to assess these patterns based on 2D pose data enabling this kind of analysis with widely available

RGB cameras. We develop an image-based VC requiring a typical laptop with a built-in webcam. The VC can provide proper exercise therapy and interact with the user to keep ones' interest in therapy.

2. Related Work

2.1. Virtual Assistive Systems for Upper Extremity Rehabilitation After Stroke

Previous works investigated computer-based solutions for in-home rehabilitation therapy with available technological devices. They tried to give answers to the main challenges regarding such systems' development - manage proper training, be affordable, suitable for home use, perform a reliable performance assessment, and interact with patients' to keep them motivated. Duff *et al.* [6] presented the AMRR, an interactive system capable of produce visual and audio feedback from the kinematic analysis of patients' movements. Brokaw *et al.* [2] introduced the HAMSTER, a Kinect game for upper limb rehabilitation, focused on compensation behaviors restriction, with a graphical interface displaying activities' representation and providing error messages with audio cues. Rikakis *et al.* [19] developed a Kinect-based system, with a tablet computer, a mat, and smart objects for reaching and grasping tasks. The tablet displays task instructions and direct performance ratings.

To assess their systems' usability and impact in real patients, researchers conducted studies with impaired subjects with light supervision [19, 2, 6]. These studies enhance the importance of systems with a simple technical infrastructure for home use and reliable motor assessment for independent use. They also highlight the relevance of proper interaction structures with visual and audio feedback and instructions, providing performance self-assessment. Patients' improved their motor function and movement quality after using the system.

2.2. Quality of Movement Assessment During Exercise Performance

With the growth of in-home rehabilitation, the need for objective metrics to track patients' progress over time enhanced the development of quantitative and automated methods to evaluate movement quality. Body joints' kinematic study is significant to describe motion patterns. This study is possible due to motion capture devices. For objective assessment to gain clinical acceptance, researchers need to prove motion capture systems and methods' reliability and feasibility.

Murphy *et al.* [14] and Ozturk *et al.* [16] identified the kinematic variables that best describe motion patterns and distinguish healthy participants from stroke survivors. Olesh *et al.* [15] and Lee *et al.* [12] provided automated methods to produce assessment scores highly correlated with FMA scores. With their work, they verified that, for the compensation component, post-stroke survivors demonstrate lower inter-joint coordination between elbow and shoulder and more severe shoulder elevation and trunk displacement patterns. The four works explored the shoulder abduction and elevation angles, and trunk displacement from its initial position to describe motor compensation. Lee *et al.* [12] explored 0, 1, and $2/5$ joints (Figure 4) projected trajectory, which is given by the distance of these joints to their initial position, at each timestamp.

Despite relevant analysis and good results from these works, they do not provide a comprehensive assessment and feedback about compensation and its different types.

3. Methodology

3.1. Virtual Coach

This section details the VC development for upper extremity rehabilitation. Introduces requirements, define the VC intelligent agent, and describe our system architecture and implementation details.

From the related work and therapists' advice, we list a set of system requirements:

- Present an exercise demonstration;
- Propose adequate exercises;
- Give patients the possibility of exercising sitting in a chair, contributing to their confidence and physical safety;
- Display of the patient's image while exercising as if looking at a mirror;
- Provide clear and repetitive audio instructions, cues for posture correction, encouragement, and suggest task repetition;
- Display visual markers indicating the arm target position and the existence of compensation.

The VC proposes three appropriate exercises (Table 3) and monitors user compensation behaviors during their execution. First, it verifies if the patient is correctly positioned to enable motion capture. Once the user is well placed, the exercise begins and it starts evaluating one's movements. It gives verbal and visual instructions about the exercise and target position the user has to reach. When the patient exhibits compensatory, the VC suggests posture improvement. It also encourages movement repetition and praises the user when the target position is reached.

State Space $\mathcal{S}$	Description
$o \rightarrow out$	Patient not placed in the correct position
$i \rightarrow in$	Patient placed in the correct position
$e \rightarrow exercise$	Exercise and movement trial beginning
$n \rightarrow normal$	Normal movement pattern
$tr \rightarrow trunk\ rotation$	Patient rotates the torso
$se \rightarrow shoulder\ elevation$	Patient elevates the shoulder
$td \rightarrow trunk\ displacement$	Patient displaces the torso
$tg \rightarrow target$	Patient reaches the target position

Table 1: states.

Rules	Description
$state_{prev} = o$ $state = o$ $time > th_{pos}$	Patient not well-positioned for $time > th_{pos}$: VC suggests body repositioning; position rectangle in red color.
$state_{prev} = o$ $state = i$	Patient well-positioned: position rectangle in green color; VC gives exercise directions.
$state_{prev} = i$ $state = e$	Exercise beginning: VC displays target position marker (green).
$state_{prev} = \mathcal{S}/\{i\}$ $state = e$	Patients stops moving: VC proposes movement repetition.
$state_{prev} = e$ $state = n$	The VC starts evaluating patient's performance and asks one to reach the target position.
$state_{prev} = \{tr, se, td, n\}$ $state = \{tr, se, td, n\}$ $time > th_{tg}$	Patient takes too much time reaching the target position: VC encourages patient to reach the target.
$state_{prev} = \{tr, se, td, n\}$ $state = tg$	Patient reaches the target: VC praises the patient; target position marker in blue color.
$state_{prev} = \{tr, se, td, n\}$ $state = tr$	Patient describes trunk rotation: VC suggests posture correction; it displays trunk compensation marker (red).
$state_{prev} = \{tr, se, td, n\}$ $state = se$	Patient describes shoulder elevation: VC suggests correction; VC displays shoulder compensation marker (red).
$state_{prev} = \{tr, se, td, n\}$ $state = td$	Patient describes displaces the torso: VC suggests posture correction; VC displays trunk compensation marker (red).

Table 2: Actions.

3.1.1. The Intelligent Agent

Our VC is a *Simple Reflex Agent*. It selects the *action* to take based on the current environment's *state*, previous state and time. Table 1 describes the different perceived states by the VC.

Table 2 describes the VC actions and trigger corresponding. VC actions are conducted through the UI. The actions include:

- Display of position markers - the rectangle indicating patient's valid positions;
- Display of the target position marker;
- Display of compensation indicator markers - shoulder and trunk markers;
- Audio speech and respective subtitles - instructions, suggestions, encouragement, and praise.

3.1.2. User Interface

To establish an interaction between the user, we developed a UI through a web application. The user chooses the training exercise and can watch each exercise demonstration. To develop our web application, we used Flask microframework [17]. Flask is written in Python. We created a dynamic web application to run locally in a personal computer with four web pages: **Init**, **Menu** (exercise selection), **Demo** (exercise demonstration), and **Main**, in which the user exercises and interects with the system.

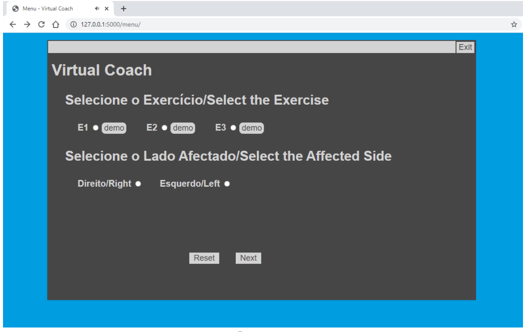


Figure 1: Virtual Coach **Menu** web page.

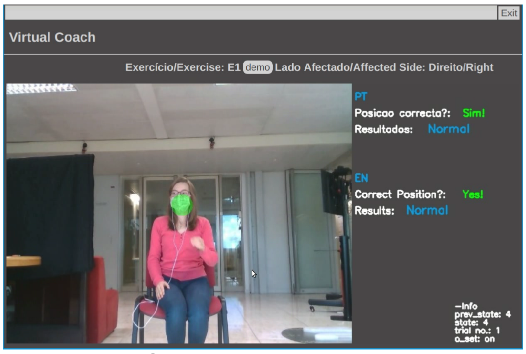


Figure 2: Virtual Coach **Main** web page - display E1 target position.

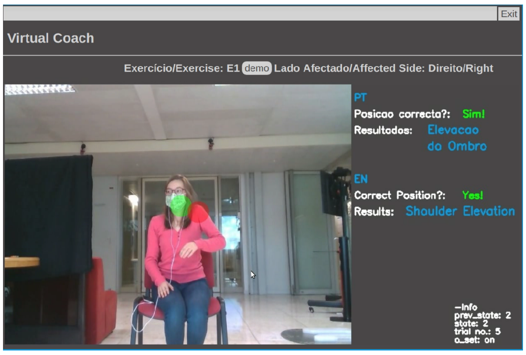


Figure 3: Shoulder elevation in E3 - display trunk compensation marker.

3.2. Compensation Assessment Methods

This section describes the proposed methods to assess compensation patterns during exercise performance from 2D body keypoints. This assessment is composed of the following steps: **Feature Extraction**, **Feature Selection**, **Data Normalization**, **Classification**, and **Result Filtering**. We propose two classification approaches - Rule-based (RB) classification approach, which works as a baseline method, and a Neural Network (NN) based approach.

3.2.1. Feature Extraction and Selection

To extract the body joints' 2D pose data, we use the OpenPose [4], a software library that provides the 2D location of 25 body keypoints in the image coordinate system, $\{I\}$. Each keypoint provided by OpenPose is denoted by $o_j^t = [p_j^t\ s_j^t = [x_j^t\ y_j^t\ s_j^t]'$, where s_j^t is a confidence score, j denotes a body joint (figure 4) from a set of joints J, and t the frame number.

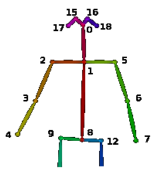


Figure 4: OpenPose Body keypoints.

While selecting the most relevant keypoints to describe patients' movements we consider three scenarios (S1, S2, and S3) concerning patients' position in front of the camera: facing the recording camera (S1), with the affected arm facing the camera in a perpendicular (S2) and oblique (S3) position. For these latter just the affected side is visible in the image. Keypoint selection follows [12]. For S1 we select the joints $J_s = [0, 10] \cup \{12\}$. For S2 and S3 $J_{s_{right}} = [0, 2] \cup [5, 8] \cup \{12\}$ $J_{s_{left}} = [0, 5] \cup \{8\} \cup \{10\}$. These keypoints are held by the RB and NN based approaches. The head keypoints, $j \in [15, 18]$, are held for the RB method, in addition to the selected joints, to overcome the lack of 3D data by head size variation.

Considering a multi-person setting (with the patient under evaluation and a caregiver), we select the patient assuming he/she is the closest person to the center of the image. Thus, given an image with a resolution of 640×480, we consider an imaginary disk in the image center defined by $D(c_d, r)$, which is a set of points in the image of the form $\{p \in [0, 640] \times [0, 480] : \|p - c_d\| = r\}$, where $c_d = [320\ 240]'$ is the disk center, $p \in [0, 640] \times [0, 480]$ and r is the disk radius. The euclidean distance between the subject and the disk center is given by $d(p_8, c_d) = \|p_8 - c_d\|$.

3.2.2. Data Normalization

In a real-world setting, subjects have body parts' of different sizes and are not placed at the same place regarding the camera. For this reason, we normalize the keypoints. Our feature normalization approach consists of three steps: transformation, normalization, mirror. First, we apply rigid body transformation from the image coordinate system, $\{I\}$, to the body coordinate system, $\{B\}$, in which the patient's joint 8 is the origin. This step considers the patients' affected side. For S1, the ${}^BX$ axis is directed to the affected side. For S2 and S3, the ${}^BX$ axis is directed to the patients' front. Additionally, we normalize each resultant keypoint coordinates to the spine length, $d^1(p_1, p_8)$, measured in $t = 1$. For the NN, to give the healthy side as a reference, we mirror the joints to the ${}^BX$ axis positive side, aligning both sides. For RB, each keypoint moves regarding other specified keypoint.

3.2.3. Kinematic Variables

For the RB classification approach we consider as features kinematic variables as in related work [12]. However, since we only have 2D positional data, we hypothesize measures to assess some compensation patterns for the three scenarios. We intend to identify four types of compensation: **trunk forward (TF)**, **trunk rotation (TR)**, **shoulder elevation (SE)**, and **other (O)** trunk compensation patterns, such as trunk moving backward and trunk tilt. Normal movements we call **normal (N)** To calculate the kinematic variables the mirror keypoint normalization step is not applied. More specifically, for scenarios S1, S2, and S3 - the formulated hypotheses and respective kinematic variables are summarized as follows.

Trunk Forward/Backward: S1 - observed changes in patient's head size, ΔH^t (H^t - head area in $t > 1$); S2 and S3 - spine angular and linear displacements, $a^t(p_8^1, p_1^1, p_1^t)$ (a^t - angle between three joints) and $\Delta x^t(p_1^t, p_1^1)$ (Δx^t - displacement in ${}^BX$).

Trunk Rotation: S1 - simultaneous angular displacements of both shoulders, $a^t(p_2^1, p_1^1, p_2^t)$ and $a^t(p_5^1, p_1^1, p_5^t)$; S2 - shoulder displacement regarding joint 1 in ${}^BX$, $\Delta x^t(p_{2/5}^t, p_1^t)$; S3 - absolute changes in the observed chest length, $|\Delta d^t(p_2^t, p_5^t)|$ (d^t - Euclidean distance between two joints).

Shoulder Elevation: S1 - shoulder elevation angle $a^t(p_{2/5}^1, p_1^1, p_{2/5}^t)$; S2 and S3 - shoulder displacement regarding joint 1 in Y, $\Delta y^t(p_{2/5}^t, p_1^t)$ (Δy^t - displacement in ${}^BY$).

Trunk Tilt: S1 - spine angular displacement $a^t(p_8^1, p_1^1, p_1^t)$; S2 and S3 - absolute changes in patient's head size, $|\Delta H^t|$.

3.2.4. Classification Approaches

Multilabel Classification (MLC) is a specific type of classification task, in which output is not a unique output value but an array of outputs [9]. We consider our problem of determine distinct compensation patterns from video frames a MLC problem. While dealing with this kind of problem we consider label dependency (when a label is active other is active too) and label imbalance (labels more frequent then others). To deal with this problem, we apply the binarization technique [10, 11] *One-vs-Rest*, which trains classifier for each label against all others [18], thus one label prediction does not influence the other. Applying this method to the original dataset generates as many predictions as the number of labels, which are then combined to produce the multilabel response.

RB classification models are a set of *if-then rules* applied to a collection of features and providing a predicted label. This kind of model has the advantage of easy comprehension and result interpretation [1, 9]. We apply a set of independent rules to the hypothesized kinematic variables in section 3.2.3. We define rules for each pattern and each scenario S1, S2, and S3, as follows:

Trunk Forward/Backward: S1 - $\Delta H^t < th_{TB}$[1] $\rightarrow P$[2] $=$ 'O', $\Delta H^t > th_{TF} \rightarrow P =$ 'TF', $P =$ 'N' $otherwise$; S2 and S3 - $a^t(p_8^1, p_1^1, p_1^t) > th_{TF} \wedge \Delta x^t(p_{2/5}^t, p_1^t) > 0 \rightarrow P =$ 'TF', $a^t(p_8^1, p_1^1, p_1^t) > th_{TF} \wedge \Delta x^t(p_{2/5}^t, p_1^t) < 0 \rightarrow P =$ 'O', $P =$ 'N' $otherwise$.

Trunk Rotation: S2 - $\Delta x^t(p_{2/5}^t, p_1^t) > th_{TR} \rightarrow P =$ 'TR', $P =$ 'N' $otherwise$; S3 - $|\Delta d^t(p_2^t, p_5^t)| > th_{TR} \rightarrow P =$ 'TR', $P =$ 'N' $otherwise$.

Shoulder Elevation: S2 and S3 - $\Delta y^t(p_{2/5}^t, p_1^t) > th_{SE} \rightarrow P =$ 'SE', $P =$ 'N' $otherwise$.

Trunk Tilt: S1 - $a^t(p_8^1, p_1^1, p_1^t) > th_{TI} \rightarrow P =$ 'O', $P =$ 'N' $otherwise$; S2 and S3 - $|\Delta H^t| > th_{TI} \rightarrow P =$ 'O', $P =$ 'N' $otherwise$.

Trunk Rotation & Shoulder Elevation S1: $(1)\ a^t(p_{2/5}^1, p_1^1, p_{2/5}^t) > th_{sh_{affected}} \wedge (2)\ a^t(p_{5/2}^1, p_1^1, p_{5/2}^t) < th_{sh_{opposite}} \rightarrow P =$ 'SE', $(1) > th_{sh_{affected}} \wedge (2) > th_{sh_{opposite}} \wedge (1) - (2) < th_{sh_{affected}} - th_{sh_{opposite}} \rightarrow P =$ 'TR', $(1) > th_{sh_{affected}} \wedge (2) > th_{sh_{opposite}} \wedge (1) - (2) > th_{sh_{affected}} - th_{sh_{opposite}} \rightarrow P =$ 'TR'&'SE', $P =$ 'N' $otherwise$.

For the NN based approach we adopt an approach to overcome label dependency issues. When we appraise **Normal** movement patterns (without compensation), we desire that our multilabel classifier is robust enough to not assign a label to a frame denoting compensation as it confirms the good movement quality. With this desire, we divide our problem into two problems, a binary and a multilabel. First, a binary classifier (C1) determine compensation existence. Second, a multilabel classifier (C2) concludes the described compensation patterns from the frames with compensation detected by C1. Figure 5 represents our proposed approach.

Figure 5: NN based approach to assess compensation.

[1] th_c - Threshold value, c denotes a compensation pattern, e.g., $c = TB$ = trunk backward.

[2] Predicted Label.

3.2.5. Result Filtering

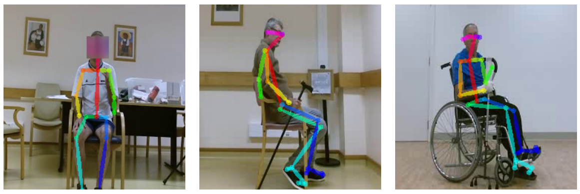
Given the classification results we filter them to produce a final decision. Classifying frames with a frame rate of $30\ fps$ is extremely exhaustive and incompatible with our human perception. To produce a final decision we establish a window of frames from which we compute its median predicted label. The median is computed for each label.

4. Experiments

This section presents all the experimental procedures conducted to evaluate the Virtual Coach (VC) and validate the compensation assessment approaches.

4.1. The Dataset

To train and validate our classification models, we use the dataset from Lee *et al.* [12] work. This dataset is of videos of a set of 15 post-stroke survivors performing three exercises. The 15 participants, with an average age of 63 ± 11.43 years old, suffered a stroke and were left with a more affected body side (left or right) [12]. Each participant performed an average of 10 trials for each exercise. The three upper limb exercises (E1, E2, and E3) are introduced in Table 4. Also, the table corresponds a scenario concerning patients positioning to each exercise.



(a) E1 - S1 **(b)** E3 - S2 **(c)** E3 - S3

Figure 6: Scenarios and Exercises.

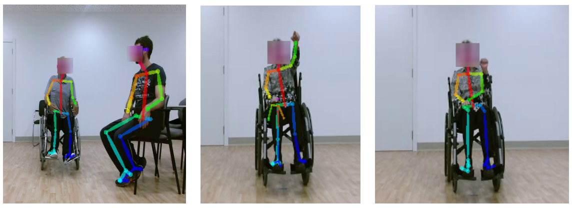
4.1.1. Labeling Process

To perform MLC, we labeled each dataset video frame. For this labeling process, we followed Physical therapist and an Occupational advices.

Acknowledging the distinct compensation patterns mentioned in section 3.2.3 we specify five labels presented in Table 4. For E1 and E2 we assigned labels *'1'*, *'2'*, *'3'*, and *'4'*. For E3, we assigned labels *'0'*, *'2'*, and *'4'*. We assigned labels to each frame in which a compensation patterns is visible even to the frames in which the behavior was only beginning or ending. In these frames the patterns are not extremely obvious. Since these set of frames correspond to movement patterns transitions we call them **borders**.

4.1.2. Data Cleansing

After keypoint extraction with OpenPose, it is important to consider three different situations: the presence of other people in the image beside the patient, extra skeletons detected, which do not necessarily belong to a person, and keypoint misdetection. Figure 7 illustrates each one.



(a) Extra person **(b)** Extra skeleton **(c)** Misdetection

Figure 7: OpenPose incorrect detection and extra person.

Extra people in the image is easily overcome with the method proposed in section 3.2.1. Extra skeletons often do not have spine joins, $\{0, 1, 8\}$, and thus since their confidence score is zero we remove them. For keypoint misdetection, we consider as relevant joints of the affected side and opposite shoulder. This joints good detection correspond to a confidence score higher then a defined value. The remaining joints must have $s_j^t > 0$. Other misdetections with a good confidence score were corrected.

4.1.3. Multilabel Dataset Characteristics

Before developing a classification model and training it, we have to know our Multilabel Dataset (MLD) characteristics concerning its multilabel nature. We use two metrics: P_{min} which is the percentage of the data samples with only one label active. A high P_{min} percentage means that most samples are single labeled, and the dataset is not greatly multilabel. As shown in Table 3, the dataset is almost single labeled - high percentage of single labeled frames, P_{min}. Regarding label imbalance, in Table 4, the $IRLbl$ metric shows the ratio between the occurrences of the most frequent label and each label. We can see that, for the three exercises, label *'4'* is the most frequent, $IRLbl = 1$. For E1 and E2, *'1'* is poorly represented, $IRLbl \gg 1$, with only one patient exhibiting this compensation pattern. For E3, the less representative label is *'2'*.

	Exercise	Scenario	P_{min}
$E1$	*'Bring a Cup to the Mouth'*	S1	83.83%
$E2$	*'Switch a Light On'*	S1	91.4%
$E3$	*'Move a Cane Forward'*	S2 & S3	98.15%

Table 3: The three exercises and percentage of single labeled frames.

	$IRLbl$		
Label	E1	E2	E3
'0: Trunk Forward'	-	-	3.54
'1: Trunk Rotation'	16.23	19.25	-
'2: Shoulder Elevation'	2.15	3.03	15.77
'3: Other'	4.93	5.55	-
'4: Normal'	1	1	1

Table 4: Considered labels and $IRLbl$ metric for each one.

4.2. Classification Methods

Once we have the data of our MLD cleaned, keypoints normalized, and we have perfect knowledge about its characteristics, we can set the threshold values for the RB method and train the NN based classifier, and validate our models. This section describes the adopted metrics to evaluate our models' performance and validation method. For the RB method, we validate the kinematic variables presented and hypothesized in section 3.2.3. For the NN, we describe the explored hyperparameters. Additionally, we describe two experiments to applyied to the obtained classification results.

We apply the classification methods to the normalized keypoints raw and filtered signal. Applying a moving average filter with a window of five frames as in [12], we reduce signal noise.

4.2.1. Evaluation Metrics

To evaluate our classification models' performance on predicting compensation patterns from video frames, we use a set of performance metrics appropriate to a MLC problem. While in the binary context, the output result from a classifier can only be considered correct or incorrect, in the multilabel field, the provided output is a set of labels (vector of 0s and 1s), being considered completely correct, partially correct, or totally incorrect [8]. This way, we need adequate metrics that acknowledge these possibilities. We use $Precision$, $Recall$, F_1 score, and $HammingLoss$ [8, 18] to evaluate our models. The first four are calculated according *micro-averaging* strategy. In *micro-averaging*, the counters of correct and incorrect predictions are joined together and then the metric is calculated, this way rare labels are diluted between the most frequent labels [8, 18].

$Precision$ is the percentage of predicted labels truly significant for the sample. $Recall$ refers to the classifier's ability to detect all positive samples. F_1 is the result of combining $Precision$ and $Recall$ metrics in an weighed harmonic mean. $HammingLoss$ is the portion of mispredicted labels.

4.2.2. Validation Method

After defining our classification methods, we need to select the best model hyperparameters that regulate our learning model's behavior and validate it. We split the MLD into a *training* set and *validation* set [7, 18].

Since we have a pretty small dataset with some labels poorly or nothing represented, we resort to *cross-validation* to evaluate our models' predictive ability and ensure generalization. Cross-validation consists of partitioning the dataset into small subsets. In the validation loop, all the sets except one are used for training, and the remaining set is used for validation [7, 18]. In the end, the performance measure determined in each loop, is averaged. Since all patients in a post-stroke status present their own motor pattern, we apply *Leave-One-Subject-Out* (LOSO) cross-validation. Validating the models on each patient compensation pattern enables a better understanding of their classification performance and generalization capacity.

4.2.3. Kinematic Variables Validation

At this stage, we validate the formulated hypotheses regarding the kinematic variables defined in section 3.2.3. This step is crucial to prove our hypotheses' efficiency to assess motor compensation from 2D data. Also, kinematic variables' analysis allows us to determine the thresholds values for the RB method. We validate this measures for the patterns observed in the dataset.

Given the rules in section 3.2.4, in the following figures we can observe that our hypotheses to assess compensation (section 3.2.3) are valide.

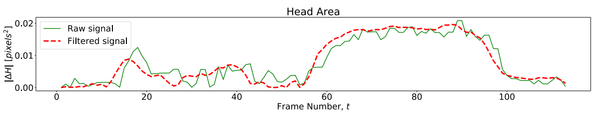


Figure 8: Head area over time of a trunk tilt (**Other**) simulation in similar conditions of the dataset for E3.

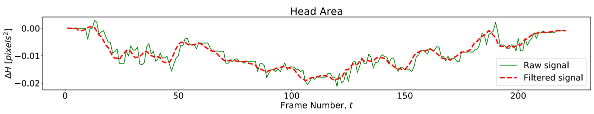


Figure 9: Head area over time, revealing trunk moving backward (**Other**) observed in the dataset for E2.

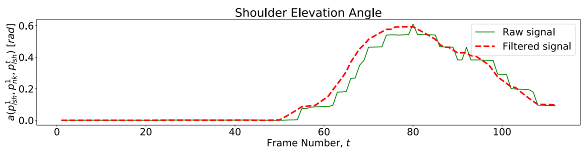


Figure 10: Patient affected shoulder elevation angle revealing **Shoulder Elevation** for E2.

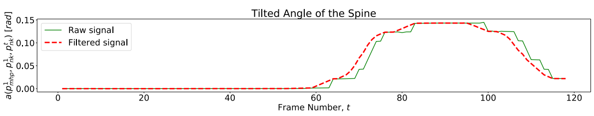


Figure 11: Patient tilted angle of the torso describing a trunk tilt (**Other**) for E2.

4.2.4. Filtering of the Classification Results

For the filtering of classification results describe in section 3.2.5, we experiment this procedure for sets of 5, 7, 9, and 11 frames. Additionally we perform another experiment with the predicted labels.

From the video frames' labeling process described in 4.1.1, another issue emerges regarding the classification results. While labeling the

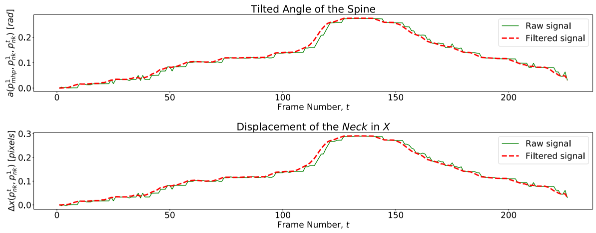


Figure 12: Patient P06 tilted and of the spine and shoulder displacement over time, describing **Trunk Forward** in E3.

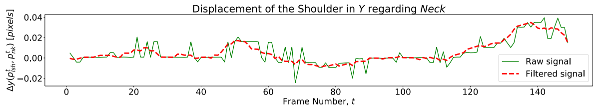


Figure 13: Patient P05 shoulder displacement over time, describing **Shoulder Elevation** in E3.

dataset, we assign labels to frames indicating compensation patterns, in which those patterns are visible. This labeling includes the frames in which the patient starts or ends describing a compensation pattern, and thus this pattern is not very pronounced. These frames are called **borders**. Given this, we assume a high probability of our classifiers reveal low performance when classifying these motion patterns transitions. we consider borders of 5 and 10 frames.

4.3. Virtual Coach

To evaluate the VC we performed a set of empirical experiments with a group of volunteers. We aim to investigate the system's **hedonic (H)** and **utilitarian (U)** value, **systems' performance (SP)**, and **users' use intention (IU)**. Thus, we introduce the following hypotheses:

H1 There is a disparity between stroke survivors volunteers and the other volunteers among the different perceptions about the system;

H2 There is a difference in the perceived Virtual Coach utilitarian value between older adults and younger adults since stroke is common among older adults and the elderly;

H3 Hedonic value perceptions are affected by the Virtual Coach performance monitoring exercise performance, detecting compensation, and interaction model awarenesses;

H4 Hedonic and utilitarian value perceptions affect users' intention to exercise with the Virtual Coach.

Data collection and storage is in agreement with the General Data Protection Regulation (GDPR). To ensure these conditions, the Instituto Superior Técnico Ethics Committee reviewed and approved our experimental protocol. For this stage we follow a previous work [3].

VID	Age	Sex	ND/A side	(a)	(b)	(c)	(d)
V01	25-34	M	Left	Y	Y	Y	Y
V02	55-64	F	Left	Y	N	Y	Y
V03	65-74	F	Left	Y	N	Y	Y
V04	65-74	M	Left	Y	N	Y	Y
V05	25-34	M	Left	Y	N	Y	N
V06	55-64	M	Left	Y	N	Y	N
V07	25-34	F	Left	Y	N	N	N

Table 5: Profiles of the volunteers. General information: (a) Knows what a stroke is (b) Had a stroke (c) Some relative or close friend had a stroke (d) Followed the rehabilitation process closely. VID - Volunteer ID; ND - Non-Dominant; A - Affected; F - Female; M - Male; Y - Yes; N - No.

4.3.1. Volunteers

We recruited seven volunteers to exercise their limbs with our system. Given the pandemic situation of the COVID-19, volunteers are among our closest social groups, such as family and close friends. When selecting the participants, we aimed to gather a diverse group concerning age, sex, and experience with the stroke thematic. Volunteers signed an Informed Consent authorizing the recording of their image necessary to the normal system operation. Table 5 present the volunteers' profiles and general information.

4.4. Experimental Setup

Motivated to provide an affordable and accessible solution with a simple technical infrastructure, we only use a laptop with a built-in webcam in this experiment. To we use the RB classification algorithm, which enables an easy result interpretation and adjust the rules' threshold values. The sessions took place in a domestic environment spacious enough to assure the laptop was placed in from the volunteer with an ideal distance to capture the participant's relevant body joints.

4.5. Experimental Procedure

At the beginning of the session, the researcher introduced the study and properly explained the entire procedure and introduced the UI to the volunteer giving an overview of its functionalities. The volunteers were asked to perform the three upper extremity exercises with their arm - from their affected side due to stroke if it was the case, or from their non-dominant body side. During the exercises, patients had to simulate the different compensation strategies mentioned in the previous sections. Volunteers repeated the movements at least five times. The session did not exceed 30 minutes. In the end, each participant answered a questionnaire giving their feedback about the VC and the interaction with it.

4.6. Data Collection

The questionnaire to volunteers provided quantitative and qualitative data. The participants responded to each question on a five point Likert scale (quantitative) - from *'1 = Strongly Disagree'* to *'5 = Strongly Agree'* - and a question with open answer for each item - e.g., ”tell me more about it” - to gather more information (qualitative).

5. Results & discussion

This section presents the experimental results. First, we evaluate the two proposed MLC approaches and analyze experiments applied to both methods' classification results. Second, we present the information collected from questionnaire respondents and a statistical analysis which validates the hypotheses raised in section 4.3. These results validate our system for autonomous upper extremity rehabilitation therapy, mainly its efficiency and relevance.

5.1. Classification Results Analysis

In this section, we compare the results obtained in the evaluation metrics among the different methods and exercises.

For the NN approach we used the hyperparameters from Table 6 for both C1 and C2 classifiers.

	C1		
	#layers	#units/layer	Learning Rate$_{init}$
E1	1	16	0.001
E2	2	16	0.001
E3	1	96	0.01
	C2		
	#layers	#units/layer	Learning Rate$_{init}$
E1	1	64	0.001
E2	1	16	0.01
E3	1	16	0.001

Table 6: NN based approach classifiers' hyperparameters.

We gather in Table 7 the results obtained for both approaches. For the RB filtered signal revealed better results, as expected. With the noise reduced, and without abrupt changes in the acquired keypoints', the application of the conditions based on a threshold value limiting the existence of compensation works better. The NN based revealed better results for raw signal.

Looking at Table 7 we can notice that for E1 RB approach performs better and for E2 and E3 the NN based approach presents better results. An evident difference between these exercises' datasets is their percentage of single labeled samples, P_{min}. For E1 has 83.83% of the samples are single labels. For E2 and E3, 91.4% and 98.15% of the samples are single labeled. This makes us believe that the RB method handles better a scenario with a not so poor multilabel nature. On the other hand, the NN based approach is more efficient for a binary problems. A particular case is E3. Although for this exercise the NN based approach performs better it has a higher value of $HammingLoss$, meaning that this approach provides more mispredictions.

From Table 8 we can infer that filtering and removing borders produces better performance met-

E	$Precision$	$Recall$	$F1$	$Hamming\ Loss$
1_{RB}	$\mathbf{0.76 \pm 0.14}$	$\mathbf{0.78 \pm 0.12}$	$\mathbf{0.77 \pm 0.12}$	$\mathbf{0.11 \pm 0.06}$
2_{RB}	0.56 ± 0.17	0.67 ± 0.17	0.60 ± 0.17	0.19 ± 0.08
3_{RB}	0.70 ± 0.27	0.71 ± 0.26	0.70 ± 0.26	$\mathbf{0.13 \pm 0.11}$
1_{NN}	0.71 ± 0.23	0.70 ± 0.25	0.70 ± 0.24	0.18 ± 0.15
2_{NN}	$\mathbf{0.73 \pm 0.21}$	$\mathbf{0.73 \pm 0.19}$	$\mathbf{0.73 \pm 0.19}$	$\mathbf{0.15 \pm 0.11}$
3_{NN}	$\mathbf{0.80 \pm 0.22}$	$\mathbf{0.80 \pm 0.21}$	$\mathbf{0.80 \pm 0.22}$	0.14 ± 0.14

Table 7: Average results and standard deviation for the Rule-based (RB) and Neural Network (NN) methods. E - Exercise.

	Filtering		**No Borders**	
	Window	$F1 - score$	Window	$F1 - score$
$E1_{RB}$	11	0.78 ± 0.13	10	0.79 ± 0.13
$E2_{RB}$	7	0.61 ± 0.17	10	0.61 ± 0.18
$E3_{RB}$	9	0.70 ± 0.27	10	0.72 ± 0.27
$E1_{NN}$	9	0.70 ± 0.23	10	0.71 ± 0.24
$E2_{NN}$	9	0.74 ± 0.20	10	0.76 ± 0.18
$E3_{NN}$	11	0.80 ± 0.22	10	0.83 ± 0.22

Table 8: Average results and standard deviation for the RB and NN based methods after filtering.

rics values. This means that if we apply the filtering to obtain a final decision, our decision is more accurate than the initial one. From border removal, we see that classifiers had difficulty classifying these regions. Without them, the metrics are not penalized.

5.2. Virtual Coach Validation Results

Table 9 presents the mean quantitative results gathered from the questionnaires. The results reveal that most volunteers enjoyed exercising with the VC, felt motivated and interested in the exercises the interaction established pleasurable. Volunteers find the system useful to help patients improve their upper extremity movement quality and valuable for when they cannot have therapists' supervision. Subjects revealed that they would confidently keep using the system to exercise. Volunteers' perception of the system's performance is that it performed really well and fulfills its purpose and they expressed that they could trust the system evaluation of their motor performance. However, some details can be improved, such as the VC response velocity and more flexibility regarding users' initial position. Qualitative information supports these results.

Following we investigate the raised hypotheses in section 4.3: **H1** Stroke survivor volunteer was more critical with the system, giving a lower mean score for **U** ($mean = 4$) and **SP** ($mean = 3$) and showing a less **IU** ($mean = 4$). However, this difference is very small. Other volunteers present a mean score for **U** value of 5.0 ± 0.0, for **IU** of 5.0 ± 0.0, and for **SP** of 4.58 ± 0.58. Concerning **H** perception, the stroke survivor and other volunteers equally enjoyed the training and interact with the system.

H2 Older adults with age over 55 years old find the system more relevant and useful (**U** $mean = 5.0 \pm 0.0$). Also, they express a clearer intention to keep using the system (**IU** $mean = 5.0 \pm 0.0$).

	H	**U**	**IU**	**SP**
H	1	0.03	1.00	0.53
U	0.03	1	1.00	0.75
IU	1.00	1.000	1	1.00
SP	0.53	0.75	1.00	1
Minimum	3.75	4.00	4.00	3.00
Maximum	5.00	5.00	5.00	5.00
Mean	4.54	4.86	4.75	4.36
SD	0.51	0.38	0.50	0.80

Table 9: Descriptive Statistics and Pearson Correlation. *SD* - Standard Deviation.

H3 & H4 Table 9 presents a summary of descriptive statistics and the Pearson correlation between **H**, **U**, **IU**, and **SP**. The Pearson correlation coefficient (ρ) measures the degree of correlation of the distinct dimensions. Table shows a correlation between **H** and **SP** with a coefficient of $\rho = 0.53$, revealing that these dimensions are linearly correlated. If the mean value of the perceived **SP** increases, it positively influences the perceived **H**, supporting H3. **IU** has a perfect linear correlation with all the other dimensions ($\rho = 1$). If the mean perceived **H** and **U** increase, **IU** grows, supporting H4.

6. Conclusions

Given the presented results, it is possible to infer that we achieved the project objectives: we were able to assess motor compensation from 2D positional data with pretty good accuracy and our VC fulfilled its purpose and met all the requirements concerning this kind of system. Given this, we formulated some improvements and future work suggestions.

To continue the investigation methods to detect motor compensation from 2D positional data, we suggest the development of a Recurrent Neural Network based method considering keypoints in sequential order. Concerning the assessment methods, our RB approach could be improved, giving priority to trunk displacements over shoulder elevation patterns, to overcome some misdetections. To evaluate the real impact of the VC in real stroke survivors under a rehabilitation process and gather more solid perception about the system.

References

[1] O. Biran and C. Cotton. Explanation and Justification in Machine Learning: A Survey. *IJCAI-17 workshop on explainable AI (XAI)*, 8(1), 2017.

[2] E. B. Brokaw, E. Eckel, and B. R. Brewer. Usability evaluation of a kinematics focused Kinect therapy program for individuals with stroke. *Technology and Health Care*, 23(2):143–151, 2015.

[3] M. Čaić, J. Avelino, D. Mahr, G. Odekerken-Schröder, and A. Bernardino. Robotic Versus Human Coaches for Active Aging: An Automated Social Presence Perspective. *International Journal of Social Robotics*, 12(4):867–882, 2020.

[4] Z. Cao, G. Hidalgo Martinez, T. Simon, S. Wei, and Y. A. Sheikh. Openpose: Realtime multi-person 2d pose estimation using part affinity fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019.

[5] T. M. Damush, L. Plue, T. Bakas, A. Schmid, and L. S. Williams. Barriers and facilitators to exercise among stroke survivors. *Rehabilitation Nursing*, 32(6), 2007.

[6] M. Duff, Y. Chen, L. Cheng, S.-m. Liu, P. Blake, S. L. Wolf, and T. Rikakis. Adaptive Mixed Reality Rehabilitation Improves Quality of Reaching Movements More Than Traditional Reaching Therapy Following Stroke. *Neurorehabilitation and Neural Repair*, 27(4):306–315, 2012.

[7] I. Goodfellow, Y. Bengio, and A. Courville. *Deep Learning*, chapter 5, pages 98–164. MIT Press, 2016. http://www.deeplearningbook.org.

[8] F. Herrera, F. Charte, A. J. Rivera, and M. J. Jesus. Case studies and metrics. In *Multilabel Classification: Problem Analysis, Metrics and Techniques*, pages 33–64, Switzerland, 2016. Springer.

[9] F. Herrera, F. Charte, A. J. Rivera, and M. J. Jesus. Introduction. In *Multilabel Classification: Problem Analysis, Metrics and Techniques*, pages 1–16, Switzerland, 2016. Springer.

[10] F. Herrera, F. Charte, A. J. Rivera, and M. J. Jesus. Multilabel classification. In *Multilabel Classification: Problem Analysis, Metrics and Techniques*, pages 17–32, Switzerland, 2016. Springer.

[11] F. Herrera, F. Charte, A. J. Rivera, and M. J. Jesus. Transformation-based classifiers. In *Multilabel Classification: Problem Analysis, Metrics and Techniques*, pages 65–80, Switzerland, 2016. Springer.

[12] M. H. Lee, D. P. Siewiorek, A. Smailagic, A. Bernadino, and S. Bermúdez I Badia. Learning to assess the quality of stroke rehabilitation exercises. In *24th International Conference on Intelligent User Interfaces (IUI'19)*, pages 218–228. Association for Computing Machinery, 2019.

[13] M. F. Levin, J. A. Kleim, and S. L. Wolf. What do motor ”recovery” and ”compensation” mean in patients following stroke? *Neurorehabilitation and Neural Repair*, 23(4), 2009.

[14] M. A. Murphy, C. Willén, and K. S. Sunnerhagen. Kinematic variables quantifying upper-extremity performance after stroke during reaching and drinking from a glass. *Neurorehabilitation and Neural Repair*, 25(1):71–80, 2011.

[15] E. V. Olesh, S. Yakovenko, and V. Gritsenko. Automated assessment of upper extremity movement impairment due to stroke. *PLoS ONE*, 9(8), 2014.

[16] A. Ozturk, A. Tartar, B. Ersoz Huseyinsinoglu, and A. H. Ertas. A clinically feasible kinematic assessment method of upper extremity motor function impairment after stroke. *Measurement: Journal of the International Measurement Confederation*, 80:207–216, 2016.

[17] Pallets. Flask Documentation, May 2020. Accessed 4-Aug.-2020.

[18] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[19] T. Rikakis, J. B. Huang, A. Kelliher, K. Kitani, S. L. Wolf, J. Choi, and S. Zilevu. Semi-automated home-based therapy for the upper extremity of stroke survivors. *ACM International Conference Proceeding Series*, pages 249–256, 2018.

[20] B. Semenko, L. Thalman, E. Ewert, R. Delorme, S. Hui, H. Flett, and N. Lavoie. An Evidence Based Occupational Therapy Toolkit for Assessment and Treatment of the Upper Extremity Post Stroke. 2015.