i-Walk Intelligent Assessment System: Activity, Mobility, Intention, Communication

Georgia Chalvatzaki*, Petros Koutras*, Antigoni Tsiami*,
Costas S. Tzafestas, and Petros Maragos

School of E.C.E., National Technical University of Athens, Greece
{gchal}@mail.ntua.gr,{pkoutras,antsiami,ktzaf,maragos}@cs.ntua.gr

Abstract We present the i-Walk system[1], a novel framework for intelligent mobility assistance applications. The proposed system is capable of automatically understanding human activity, assessing mobility and rehabilitation progress, recognizing human intentions and communicating with the patients by giving meaningful feedback. To this end, multiple sensors, i.e. cameras, microphones, lasers, provide multimodal data in order to allow for user monitoring, while state-of-the-art and beyond algorithms have been developed and integrated into the system to enable recognition, interaction and assessment. More specifically, i-Walk performs in real-time and consists of four main sub-modules that interact automatically to provide speech understanding, activity recognition, mobility analysis and multimodal communication for seamless HRI. The i-Walk assessment system is evaluated on a database of healthy subjects and patients, who participated in carefully designed experimental scenarios that cover essential needs of rehabilitation. The presented results highlight the efficacy of the proposed framework to endow personal assistants with intelligence.

Keywords: intelligent assessment system, human-robot interaction, activity recognition, 3D pose estimation, speech understanding, gait tracking, gait stability, multimodal communication

1 Introduction

The rapid increase of people with special needs, such as the elderly population, and the simultaneous reduction of personal care staff, reinforce the need for robotic assistants [29]. When designing an intelligent assistant platform for people with mobility and/or cognitive impairment, special care should be given in developing a system that will monitor and promote rehabilitation in a natural and seamless way.

In order for an intelligent robotic assistant to achieve these goals, the development of advanced Human-Robot Interaction (HRI) components and their integration under a unified autonomous system is more than essential. More specifically, the platform should be capable of understanding human activities, intentions and needs, but also

* First three authors have equal contribution.

[1] This research has been co-financed by the European Union and Greek national funds through the Operational Program Competitiveness, Entrepreneurship and Innovation, under the call RESEARCH CREATE INOVATE (project code: T1EDK- 01248, acronym: i-Walk).

Proc. European Conference on Computer Vision Workshops (ECCVW), August 2020.

Eighth Int'l Workshop on Assistive Computer Vision and Robotics (ACVR-2020), Glasgow, UK, Aug.28, 2020.

Figure 1: Left: A patient walking supported by the i-Walk assistant platform. Right: A patient performing rehabilitation exercises while being monitored by i-Walk.

analysing multi-sensory signals related to gait and postural stability, so as to provide support and communication.

The i-Walk platform (Fig. 1) has been carefully designed to fill this need, by combining multisensory streams to perform a multitask understanding of human behavior, i.e., speech intention recognition, generalized human activity recognition, and mobility analysis. The multimodal interaction framework of i-Walk aims to provide natural communication and valuable feedback to the user and the medical experts regarding rehabilitation progress, in a way close to that of a personal carer.

Design and development of personal robotic assistants for elderly is prominent in scientific research [17, 27]. The role of personal care robots is multiple, covering physical, sensorial and cognitive assistance [22], health and behavior monitoring and companionship [39]. Most intelligent assistive platform designs aim to solve only specific problems, e.g. GUIDO [36] and iWalker [23] provided navigation assistance. Considerable amount of research has focused on analysing anthropometric data from various sensors for assessing human state [19] and eventually control a robotic platform, like CAIROW [13]. The ISR-AIWALKER employed RGB-D data for monitoring the users [31]. MOBOT [29] was equipped with various sensors, attempting to model human activities from multimodal data [20, 35] and perform gait and stability analysis [9, 12]. In [11] a method was proposed that integrated a human motion intention model, exploiting RGB-D and laser data of the user, into a decision making framework for adapting the platform's motion.

Several works aim to integrate visual perception into assistive robotic applications [24, 41]. Robotic assistive vision is an important topic for HRI systems [18], while also multi-sensory systems integrating visual activity with speech recognition [15] for communicating with robots have recently emerged [43,44]. Human activity recognition is a long-term research problem [47] where human skeleton representations from single images [14] have been extensively used [46, 48–50]. Deep Learning progress led to efficient methods for human pose estimation [8] while action recognition benefits from

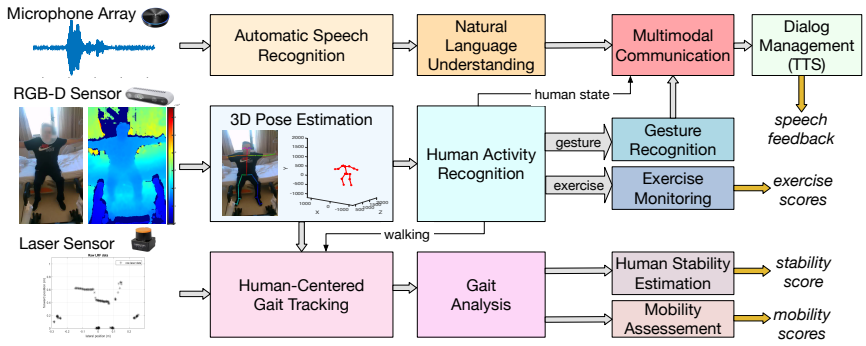

Figure 2: Overview of the multimodal i-Walk intelligent assessment system.

these improved skeleton estimations by employing recurrent methods like Long Short Term Memory (LSTM) networks, which have the ability to model temporal information of action sequences [16, 26, 38, 40, 45, 51].

However, those methods present difficulties in getting integrated in robotics, since most of them rely on datasets from constrained environments with static cameras, which makes them vulnerable to camera motion effects. Moreover, the most recent implementations incorporating human pose estimation and/or action recognition for robotic applications [33, 34, 37, 52] usually focus on specific tasks rather than provide a holistic approach that demands multimodal human perception and real-time feedback from the system.

In this paper we present the i-Walk intelligent assessment system, a multi-sensory, multimodal HRI framework destined to provide simple rollators with intelligence, aiming to be of use by patients with mobility and/or cognitive impairment. The system has three important goals: monitoring the patient in terms of his/her activity and mobility, interacting with the patient by allowing him/her to communicate his/her intentions to the platform, engaging in dialog with him/her, giving feedback, etc., and assess the patient's status in the context of a rehabilitation procedure. To this end our main contributions are the design, development, integration and evaluation of the assessment system with its individual interacting sub-modules: the speech understanding, the activity recognition, the mobility analysis and the multimodal communication system. All sub-modules communicate automatically, providing direct feedback. The proposed system has been extensively evaluated using a large corpus of multi-sensory data, both from healthy subjects and real patients from a rehabilitation center, who participated in experimental scenarios that meet the real needs of daily living and mobility rehabilitation.

2 System Overview

The i-Walk Assessment System with its respective sub-modules is presented in Fig.2. The upper flow (orange blocks) represents the *Speech Understanding* module, the middle flow (light blue blocks) the *Activity Recognition*, while the lower one (light pink blocks) the *Mobility Analysis* module. Each module exerts certain outputs for assessing the

activity state and performance, the spoken and gestural intentions and the mobility performance of the user. The *Multimodal Communication* module (red block) is responsible for triggering the dialog management (light green block) providing speech feedback to the users. These modules along with their respective parts are described below.

2.1 Speech Understanding

Speech is the most natural and instant means of communication. Thus, a speech understanding module that will enable the patient to communicate his/her intentions to the robot is a key component of the system and involves two sub-modules, as depicted in Fig. 2: An Automatic Speech Recognition (ASR) module and a Natural Language Understanding (NLU) one. For ASR, a state-of-the-art system has been integrated, where speech recorded through a microphone array serves as input to Google speech-to-text API [2] and is transformed into text. Subsequently, the transcribed text serves as input to the NLU module, in order to be translated into a human intention. The integrated NLU system has been built with RASA [3, 4, 7]: A set of pre-defined intentions, both general purpose and specific to the current application has been designed. The former category includes 7 general intents, namely greeting, saying my name, saying goodbye, thanking, affirming, denying, asking to repeat, while the latter one includes 7 intents designed for the HRI: standing up, sitting down, walking, stopping, ending interaction, going to the bathroom, doing exercises. Each intention is associated with various phrases to express this particular intention. For example, a user can express his/her will to stand up by saying "I want to stand up", "Can you please help me stand up", or any other variation. A RASA NLU pipeline called tensorflow embeddings [3] is then employed to predict the current intention based on the speech transcription. The predicted intent is the input to the multimodal communication module, that is described later in this section.

2.2 Activity Recognition

Apart from speech, gestures and human activities convey crucial information about a person's intent or state as well. We have designed and implemented a novel sub-module for human activity and gesture recognition that consists of two different subsystems: The first one performs 3D human pose estimation (Fig. 2) using the RGB-D sensor, which is mounted on the the robotic rollator (Fig. 1), while the second one recognizes human activity by employing a LSTM-based network architecture (Fig 3). In case the recognized activity is an exercise, the exercise monitoring module presents (on a screen placed on the rollator) and stores the corresponding recognition scores, while in case a gesture is detected, the gesture recognition module is triggered.

3D Pose Estimation: For the detection of the 2D body keypoints on the image plane we employ Open Pose Library [8] with the accompanied models trained on large annotated datasets [6, 25]. The third dimension of the 3D body keypoints is obtained by the corresponding depth maps. Subsequently, given a pair of pixel coordinates for a body joint and the depth value at this pixel, we calculate the corresponding 3D joint's coordinates through the inverse perspective mapping using the calibration matrix of the camera. For the final human skeleton we discard the keypoints of face, hands and feet either because in many cases they are not detected, or the depth values at these points are not reliable.

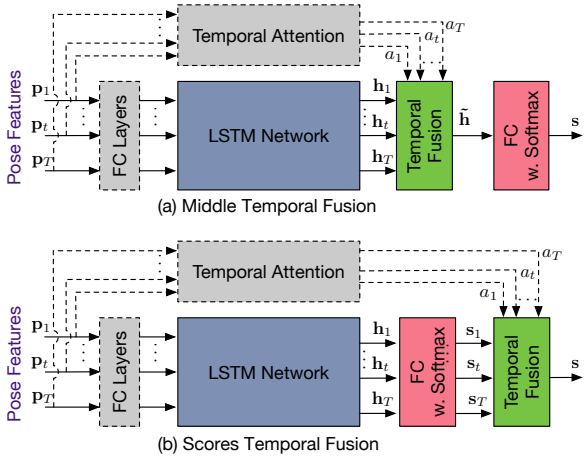

Figure 3: Neural Network architectures for human activity recognition based on LSTM units. The temporal fusion of the predictions can be applied in two different stages.

For activity recognition, either the 2D keypoints on the image plane or the 3D locations of the human joints are used as features. In the former case, since locations are pixel coordinates, we apply a standardization scheme (STD) in order to have features with zero mean and unit variance for each instance sequence. In the latter case, we transform the 3D body joint locations which are provided in the camera coordinate system, to the body coordinate system with the middle-hip joint as origin and normalized by the length between the left- and right-hip joints (BNORM scheme). In addition, we can optionally enhance the pose feature vector with the 3D velocity and acceleration of each joint, computed from the sequence of the normalized 3D joints' positions.

LSTM-based Network for Activity Recognition: In our deep learning based module for human activity recognition we employ a Neural Network architecture based on LSTM units [21]. LSTM constitutes a special kind of recurrent neural networks that can effectively learn long-term dependencies that exist in sequential data, such as human joint trajectories. Our network architecture consists of two LSTM layers stacked on top of each other (Fig. 3 - blue boxes) and a fully connected (FC) layer, followed by softmax activation, to obtain per-class scores (Fig. 3 - red boxes). The sequence of the pose features $\mathbf{p}_t$ in a temporal window of length T, possibly transformed by a sequence of FC layers, is used as input to the above network. The usage of FC layers acts as a static transformation of the initial pose features independently of time dependencies that are modelled by the LSTM units. The network output consists of a sequence of per-class labels, one for each kind of human activity.

Usually, an input sequence of length T is classified by choosing the class with the highest score $\mathbf{s}_T$ that corresponds to the the hidden state $\mathbf{h}_T$ of the last frame. However, in this work we have investigated two different approaches for the temporal fusion of the network's predictions. In the first (middle fusion: Fig. 3.a) we fuse the hidden states $\mathbf{h}_t$ according the aggregation function $\tilde{\mathbf{h}} = \mathscr{F}(\mathbf{h}_t)$, while in the second (scores fusion: Fig. 3.a) we apply temporal pooling of the softmax scores $\mathbf{s}_t$: $\mathbf{s} = \mathscr{F}(\mathbf{s}_t)$. For each one of

the above fusion schemes we have experimented with four different types of aggregation functions $\mathscr{F} : \mathbf{g} = \mathscr{F}(\mathbf{f})$, where $\mathbf{f}$ can be either the states $\mathbf{h}_t$ or the scores $\mathbf{s}_t$.
Average pooling: In this approach, $\mathscr{F}$ represents the frame predictions average, inside a temporal window: $\mathbf{g} = \frac{1}{T}\sum_{t=1}^{T}\mathbf{f}_t$. While average pooling function is able to capture information from the whole temporal video segment, it is sensitive to noisy background.
Max pooling: In this function we apply max pooling to the elements of $\mathbf{f}$ among the T frames of the temporal window: $\mathbf{g} = \mathscr{F}(\mathbf{f}) = \max_{t\in 1,\cdots,T}\mathbf{f}_t$. With max pooling we emphasize the most discriminative frame of the video segment, which has the strongest activation, and ignore the other parts that may contain noisy information.
Weighted average: An alternative way of temporal pooling is to use weights γ_t for each frame prediction, learned during the network training: $\mathbf{g} = \sum_{t=1}^{T}\gamma_t \cdot \mathbf{f}_t$. This way, we learn the relative importance of each temporal part but the weights are data independent since they are learned from the whole dataset and are not affected by the content of each video clip.
Attention weighting: To deal with the above limitations we can apply a weighting scheme where the importance weights a_t for each frame depend on the content of the video: $\mathbf{g} = \sum_{t=1}^{T} a_t(\mathbf{p}_t)\cdot\mathbf{f}_t$. More specifically, the weights are learned using an attention mechanism based on the pose features $\mathbf{p}_t$. In the first phase, the features $\mathbf{p}_t$ of each frame are transformed to attention activation values b_t using a fully connected layer: $b_t = \mathrm{FC}_{\mathrm{att}}(\mathbf{p}_t)$. Then, we compute the weights a_t by applying a temporal softmax normalization:

$$a_t = \frac{\exp(b_t)}{\sum_{\tau=1}^{T}\exp(b_\tau)}. \tag{1}$$

Network training: We trained from scratch the proposed network using mini-batches of 256 clips for 500 epochs, with initial learning rate 0.1, momentum 0.9 and weight decay 10^{-5}. The learning rate is divided by 10 after 100 epochs. For training we employed the Stochastic Gradient Descend (SGD) optimizer with the a weighted cross-entropy loss while we have augmented the corpus by flipping the original data in the vertical dimension:

$$\mathscr{L} = -\frac{1}{N}\sum_{j}^{N}\mathbf{w}_{c_j}\log\mathbf{s}^{c_j}(\mathbf{p}^{j}_{1:t},\mathbf{W}), \tag{2}$$

where c_j denotes the class of the j-th sample in the minibatch, $\mathbf{W}$ the trainable parameters of the network and $\mathbf{w}$ is a vector containing the weights for each class, based on the appearance frequencies of each class in the dataset.

2.3 Mobility Analysis

Gait stability and mobility assessment are important for evaluating the rehabilitation progress. The Mobility Analysis module, triggered when activity ''Walking" is recognized, consists of the following sub-systems (Fig. 2):
Human-centered Gait Tracking & Gait Analysis: The tracking module exploits the RGB-D data capturing the upper-body and the laser data detecting the legs. An hierarchical tracking filter based on an Unscented Kalman Filter estimates the positions and velocities of the human Center-of-Mass (CoM), which is computed by the estimated 3D human pose, while an Interacting Multiple Model Particle Filter performs the

gait tracking and the recognition of the gait phases at each time frame [9, 12]. Considering gait analysis literature [32], the walking periods are segmented into distinct strides given the gait phases recognition and certain gait parameters are computed [10, 28].
Gait Stability Assessment: A deep neural network was designed and evaluated in [9], as an encoder-decoder sequence to sequence model based on LSTMs. The input features are the estimated positions of the CoM and the legs along with the respective gait phase at each time frame, while the output predicts the gait stability state considering two classes: stable walking and risk-of-fall state. In particular, the stability score used here is the probability of performing stable walking. If this probability is below a specific threshold (defined by the the experts usually at around 30% for the most patients) the rollator's screen flashes red in order to inform the user to improve his/her gait.
Mobility Assessment: For assessing the patient's mobility status, we compute gait parameters, such as stride length, gait speed, etc., which serve as a feature vector for an SVM classifier [10]. The classes are associated with the Performance Oriented Mobility Assessment (POMA) [42]. POMA scores less than 18 refer to high risk of fall, while a score between 19 and 23 indicates a moderate risk. In this work, the mobility score is taken as the probability for a patient to belong in the high risk mobility class.

The stability and mobility assessment scores, along with several other walking parameters, like gait speed, swing phase, etc., are stored for being reported to the experts responsible for the patients' rehabilitation, so they can acquire a day-by-day quantitative information about the patient's progress.

2.4 Multimodal Communication and Feedback

As depicted in Fig. 2, the multimodal communication module is responsible for gathering and combining the outputs of the speech understanding and the gesture recognition modules and the human state (i.e. standing, sitting, etc.) so as to produce feedback for the user. It should be noted (see Fig. 4) that it can have either both inputs (speech, gesture) at the same time or asynchronous, or a single one [44]. The feedback to the user is given via a TTS (text-to-speech) system [44].

3 Experimental Results

In this section, we present the data collection scenarios and procedure, as well as the performance evaluation of the several components of the system.

3.1 Data Collection & Evaluation Setup

For the training and evaluation of the proposed multimodal assessment system, several sessions of data collection took place, involving both healthy subjects and patients with various mobility and/or cognitive inabilities. The experimental platform depicted in Fig. 1, was equipped with a RealSense camera, a Hokuyo UST-LX10 laser, and an eMeet microphone, which provided the multi-sensory data. The communication of the multiple sensors and sub-modules of the proposed system (Fig. 2) was implemented via ROS [5], while the platform was also equipped with NVIDIA Jetson TX2 modules that

Figure 4: Examples of the multimodal communication system that combines the speech understanding and the gesture recognition outputs. Note that in most cases the user's intention is successfully recognized in both modalities (green: correctly recognized intention, red: intent not recognized).

allow us to have deep learning processing capability on the rollator. In this work, we present two databases (DB) regarding the collected multimodal data:

1. i-Walk DB: This DB incorporates data both from healthy users and patients. The data collection with patients took part in DIAPLASIS Rehabilitation center [1] in Kalamata, Greece in July 2019. Thirteen patients have participated in the experiments, after approval by the medical staff. The demographic data (age and gender) along with the respective Mini-Mental Mean Score (MMSE) indicating cognitive capacity and the POMA scores regarding mobility efficiency are showcased in Table 6. All patients have signed written consent and are protected under GDPR. For development and comparison purposes, we conducted additional data collection experiments with twenty healthy users (ages 23-32). The data collection scenarios have been carefully designed in collaboration with medical experts, covering real needs regarding ambulation and rehabilitation of the patients.

Human Activities		
ID	Codename	Description
1	Sitted	sitted on chair or bed
2	StandUpPrep	preparing to stand up using rollator
3	StandUp	standing up from sitted position
4	SitDown	sitting down from standing position
5	Walking	walking using rollator
6	Standing-still	standing still without make any action
7	HandCross	place hands crossed on torso while sitted
8	HandCrossTurn	turning torso left/right with hands crossed while sitted
9	HandOpenTurn	turning torso left/right with hands opened while sitted
10	HandOpen	raise hands horizontally while sitted
11	WeightMoves	body weight transfers from one leg to another while standing supported by rollator
12	StepsHigh	in-place steps with high knees supported by rollator
13	TurnStanding	turning torso left/right while standing supported by rollator
14	Gesture	performing gesture towards rollator

Gestures		
ID	Codename	Description
a	ComeCloser	ask rollator to come closer
b	WantStandUp	want to stand up
c	WantSitDown	want to sit up
d	Stop	stop the procedure
e	End	procedure completed

Table 1: Description of the activities and gestures classes.

Target Class \ Predicted Class	1	2	3	4	5	6	7	8	9	10	11	12	13	14
1	89.8%	2.5%	1.6%	1.0%	0.0%	0.0%	0.0%	1.2%	0.0%	0.0%	0.0%	0.0%	0.0%	3.9%
2	7.1%	60.7%	0.0%	0.0%	0.0%	14.3%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	17.9%
3	0.0%	0.0%	83.7%	0.0%	0.0%	3.5%	0.0%	0.0%	0.0%	0.0%	1.9%	0.0%	0.0%	10.9%
4	1.9%	0.0%	5.1%	76.4%	0.0%	5.6%	0.0%	0.0%	0.0%	0.0%	0.0%	3.8%	0.0%	7.1%
5	0.0%	0.0%	0.0%	0.0%	88.5%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	11.5%
6	0.0%	0.0%	1.7%	3.3%	0.9%	66.0%	0.0%	0.0%	0.0%	0.0%	5.4%	3.2%	0.6%	18.8%
7	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	81.8%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	18.2%
8	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	100.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
9	3.8%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	1.5%	91.3%	0.8%	0.0%	0.0%	0.0%	2.6%
10	4.6%	7.4%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	87.9%	0.0%	0.0%	0.0%	0.0%
11	0.0%	0.0%	0.0%	1.1%	2.6%	23.8%	0.0%	0.0%	0.0%	0.0%	43.8%	10.0%	3.1%	15.6%
12	0.0%	0.0%	3.2%	15.7%	1.0%	23.4%	0.0%	0.0%	0.0%	0.0%	0.9%	52.1%	0.0%	3.8%
13	0.0%	0.0%	0.0%	1.7%	0.0%	36.7%	0.0%	0.0%	1.7%	1.7%	5.8%	0.4%	32.6%	19.5%
14	1.7%	6.2%	0.7%	2.2%	0.5%	7.3%	0.0%	0.0%	0.0%	0.2%	0.0%	0.0%	0.0%	81.1%

Table 2: Confusion matrix for the activity recognition module (middle fusion using MAX) in the i-Walk DB (patients activities).

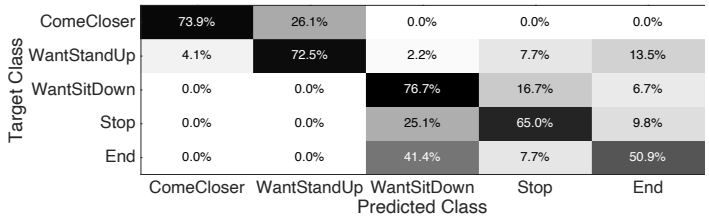

Table 3: Confusion matrix for the gestures recognition module (middle fusion using MAX) in the i-Walk DB (patients gestures).

Scenario 1: Rehabilitation exercises. The users are seated on a bed/ chair and are asked to perform certain exercises, part of a rehabilitation program. Such exercises include hands raises, torso turns, sit-to-stand transfers, etc. The complete list of the performed human activities is depicted in Table 1. The users were also prompted to perform other meaningful activities, like gestures, and express their respective intentions in free speech (Table 1).

Scenario 2: Transfer to bathroom. This scenario is essential to patients with mobility problems, as it aims to assist them in a fundamental daily-living need. The users are initially seated on a bed/ chair, and express the intend of standing up and going to the bathroom. The walking includes navigating through the hospital room, entering the bathroom, sitting on the toilet, standing up, return to the bed/chair.

2. MOBOT DB: This DB was collected in the context of the EU project MOBOT [30] in Agaplesion Bethanien Hospital in Heidelberg Germany, in 2014. This DB includes multimodal data from 14 patients who performed a minimal set of activities, namely the activities with ID: 1,3,4,5,6 listed in Table 1. We aim to showcase the transferability of the proposed activity recognition model across different setups and sets of activities.

3.2 Evaluation Results

Speech Intent Recognition: In order to assess the speech understanding module performance, we evaluate the percentage of the correctly recognized intents from speech. For this purpose, we employ the aforementioned i-Walk DB. The collected data contain free speech uttered by both healthy subjects and patients expressing a specific intent. We evaluate the performance of the intention recognition. For the healthy subjects, the DB contains 445 utterances and the accuracy is 94.83%, which is relatively high. For the patients, results are presented in Table 6, in the Speech Intent Recognition column, both for each patient separately and in average. The performance for 173 utterances reaches 74.25%, which is lower than the healthy ones', but this can be attributed to two factors: First, most of the patients did not only have mobility issues, but also cognitive/mental ones, as can be seen in Table 6, consequently their speech was often unintelligible and thus difficult to be transcribed correctly into text. Second, since the patients' age is significantly higher than the healthy ones', their utterances for the same intents are different and have larger variability. Overall, the performance is satisfying, but it could be improved by collecting more data or creating more specific acoustic models.

2 LSTM layers (without any FC layers before the LSTMs) with average score fusion						
Pose Features	2D	2D-STD	3D	3D-BNORM	3D-BNORM w. veloc.	3D-BNORM w. veloc. & accel.
MOBOT DB.	50.38	77.62	59.65	86.15	**87.72**	86.80

3D-BNORM pose features and the average score fusion scheme					
Arch.	LSTM	LSTM	FC1+LSTM	FC1+FC2+LSTM	FC1+FC2
FCs sizes	-	-	512	[1024, 512]	[1024, 512]
LSTM Layers	1	2	2	2	0
Hidden sizes	256	[256, 256]	[256, 256]	[256, 256]	-
MOBOT DB	85.47	**86.15**	85.08	83.42	71.92

Table 4: Ablation study for the activity recognition system w.r.t. different employed pose features (top) and different network architectures (bottom).

Method	Last Hidden	Middle Fusion				Score Fusion			
Aggreg. Func.	-	AVG	MAX	Weighted	Attention	AVG	MAX	Weighted	Attention
MOBOT DB (patients, 5 activities)	85.83	86.21	**90.36**	75.55	86.89	87.72	87.91	80.61	87.56
i-Walk DB (healthy users, 14 activities)	92.74	90.87	94.59	91.86	90.44	90.36	**95.20**	91.60	90.86
i-Walk DB (healthy users, 5 gestures)	80.07	67.58	88.11	78.96	68.54	70.89	**88.48**	82.89	71.30
i-Walk DB (patients, 14 activities)	68.90	72.01	73.99	72.96	69.53	68.28	70.76	**75.82**	65.45
i-Walk DB (patients, 5 gestures)	64.42	51.63	**67.81**	61.72	55.11	56.86	66.21	58.88	51.91

Table 5: Evaluation results for the different temporal fusion approaches. As pose features we have used 3D-BNORM with velocities while the network architecture consist of 2 LSTM layers without any FC layers before the LSTMs. Bold fonts stand for the best performance while the underlined fonts denote the second best.

Activity Recognition: We have conducted a series of experiments to verify our design choices in the proposed system, and investigated several variants in order to achieve the best performance, which is crucial in robotics applications.

Pose features selection and network architecture: The ablation study (following a leave-one-out cross validation approach) is presented in Table 4, using the MOBOT DB (862 clips in total) that contains simple actions and is thus suitable for running exploratory experiments regarding several parameters. We observe that the 3D features with the body normalization scheme (BNORM) outperform the 2D features, and perform even better when enhanced with the 3D velocities. We also observe that the 2-layer LSTM network without FC layers achieves the best performance. Note that the network containing only FC layers has significantly lower accuracy since it does not count for the temporal information that is necessary for recognizing dynamic activities.

Temporal fusion schemes: In Table 5 we present the evaluation results for the different fusion levels and the employed aggregation function using the features and network architecture that have achieved the best performance in the previous ablation analysis. These experiments have been conducted on the new i-Walk DB, which contains 4 different subsets, in order to validate the generalization of the proposed fusion schemes in more challenging cases. For the healthy users we conduct experiments by employing a leave-one-out cross validation strategy resulting in training and testing sets of 1891 and 107 clips respectively. For the patients we employ the whole corpus of the healthy users as training set (1998 clips) while for the testing we use patients data (1072 clips in total).

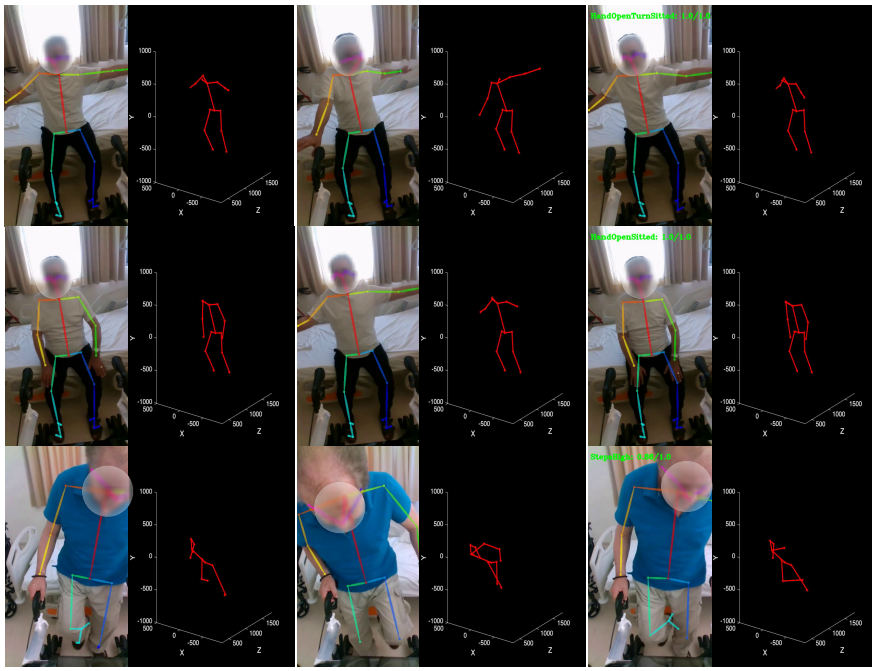

Figure 5: Examples of the activity recognition system. The system monitors the patients exercises and outputs the scores of their performance (green: correctly performed exercise).

Results indicate that the best performance overall is achieved for fusion using the MAX function, since this scheme can detect the most discriminative part of the video clip and ignore the other parts, thus recognizing better highly confusing activity classes. We can also observe that "Weighting" and "Attention" schemes perform in many cases quite higher than simple average. Regarding the different fusion levels, middle fusion with max-pooling has the best performance across the different datasets, users and action types, since it has at least the second best performance in all cases.

Evaluation analysis: Regarding i-Walk DB results, the patients' performance is lower compared to the healthy users', for both activities and gestures subsets. Moreover, gesture recognition accuracy is lower than the activity recognition one, since patients had a large variability in the way they performed gestures compared to healthy users. The confusion matrix presented in Table 2, indicates that activities performed in standing position are often confused with "still-standing" position (due to many patients' difficulty in executing actions while standing). For the same reason, gestures performed in standing position achieve also lower recognition rates (Table 3). Moreover, activities with small duration variation in pose (i.e., "StandUpPrep" or "TurnStanding") are classified sometimes as gestures. For the per patient performance (Table 6) we note that users with high MMSE achieve quite high accuracy for both activities and gestures (see examples in Fig 5. Combining gesture recognition results with speech intention recognition ones, we can observe that in some cases users with low speech intention performance achieve

Patients	Age	Gender	MMSE	POMA	Activity Performance (%)	Gesture Recognition (%)	Speech Intent Recognition (%)	Gait Speed (cm/sec) (mean±std)	Stability Score (mean±std) (%) ↑	Mobility Score (mean±std) (%) ↓
1	80	M	29/30	18/28	75.29	90.00	80.00	23.17 ± 8.36	41.93 ± 31.92	62.48 ± 30.35
2	86	M	27/30	18/28	87.01	92.31	77.77	24.83 ± 9.24	46.59 ± 34.86	63.34 ± 27.71
3	25	F	29/30	18/28	83.54	73.33	100.00	25.09 ± 8.16	40.44 ± 32.83	60.68 ± 27.59
4	83	F	23/30	11/28	80.56	75.00	78.57	21.43 ± 8.62	45.57 ± 32.42	64.54 ± 30.49
5	84	F	17/30	11/28	64.67	54.55	62.50	20.09 ± 8.74	42.64 ± 33.78	63.39 ± 27.76
6	50	M	29/30	13/28	79.75	73.33	93.33	15.54 ± 8.37	45.16 ± 34.73	61.79 ± 27.06
7	78	M	27/30	15/28	83.58	40.00	54.54	19.04 ± 9.43	50.77 ± 31.27	57.30 ± 32.31
8	73	M	18/30	12/28	52.78	52.94	86.67	21.98 ± 8.68	47.23 ± 32.37	62.14 ± 30.62
9	72	F	19/30	11/28	59.49	35.71	46.15	18.12 ± 8.62	47.92 ± 36.88	67.47 ± 28.91
10	75	M	25/30	16/28	59.77	75.00	56.25	18.96 ± 9.60	44.79 ± 33.58	65.20 ± 29.02
11	85	F	19/30	14/28	66.23	56.25	84.62	19.67 ± 9.60	47.62 ± 31.72	66.07 ± 27.03
12	55	F	28/30	16/26	90, 79	70.59	100.00	19.98 ± 9.95	31.93 ± 31.72	64.19 ± 28.67
13	75	M	28/30	11/28	91.67	91.91	46.15	26.55 ± 11.71	41.31 ± 33.33	66.24 ± 29.02
Average	-	-	-	-	75.01	67.69	74.25	21.24 ± 9.97	45.10 ± 33.86	63.80 ± 28.95

Table 6: Demographics and evaluation scores for the thirteen patients of the i-Walk database.

quite high rates in gesture recognition, a fact that highlights gestures as an alternative way for communication.

Mobility Analysis: The individual components of the mobility analysis module are adopted from works in [9, 11, 12] proving tracking robustness [9, 11] and high performance scores in stability and mobility status recognition [9, 10], hence are suitable for the multimodal setting of i-Walk framework. Building on this mobility analysis system, we fine-tuned the models with walking data of some trials of users in the i-Walk DB (from a different walking scenario not included in this work), in order to provide the necessary mobility assessment scores for monitoring rehabilitation. In particular, we evaluate the stability performance, the mobility classification, and present the gait speed parameter for each patient w.r.t. their categorization by the medical experts.

The average stability scores of each patient along with the respective standard deviations (std) are depicted in Fig. 7. The solid red line represents the average stability score of the healthy subjects and the dotted lines the upper and lower confidence levels of the healthy stability measure. It is evident that all patients present low stability while walking, and only some of them can achieve instances of stability close to the lower bound of the healthy performance. This can also be affirmed by the results in Table 6 (Stability score), where the average stability score across all patients is 45.1%, while the healthy score is 81.46%. The upward arrow means that higher scores correspond to more stable performance.

Table 6 also presents the Mobility scores and the mean and std for gait speed. All patients have been classified to the high risk-of-fall class, which is also confirmed by the POMA scores. Here, the downward arrow denotes that lower scores refer to better performance. It is interesting that patients with higher POMA perform slightly better, e.g. patient #3 (POMA 18) had average gait speed 25cm/sec w.r.t #6 (POMA 13) with average speed 15.5cm/sec. Although, this is not the norm, as more gait parameters are important for mobility classification [10], it is an indication for examining interclass categorization, for which a larger database from patients, presenting higher variation in terms of mobility, should be collected. Figure 6 presents snapshots from patients #4 and #7 with a depiction of the center of mass and legs' state estimation, that feed the stability and mobility analysis classifiers. In the current setting patient #4 is performing less stable walking than patient #7 (scores 44% against 70% of stability). Such a detec-

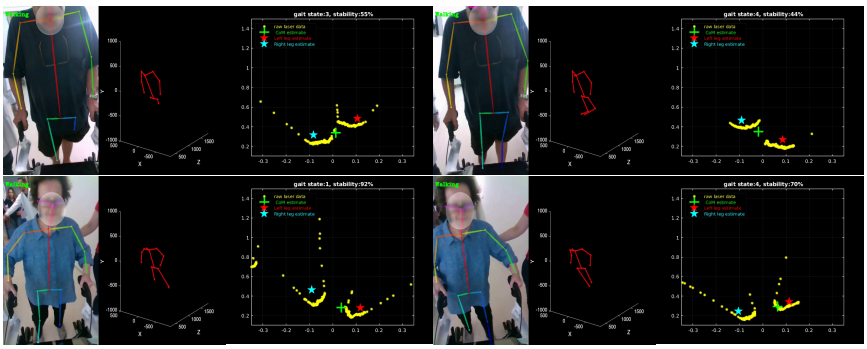

Figure 6: Mobility analysis examples. The center of mass from the detected pose is fused with the legs' state estimation for an accurate patient tracking. These observations feed the gait stability network and the mobility assessment classifier. Upper: Snapshots from patient #7 with current estimate of stability at 44%. Lower: Snapshots of patient #4 with current estimate of stability at 70% (higher score is better).

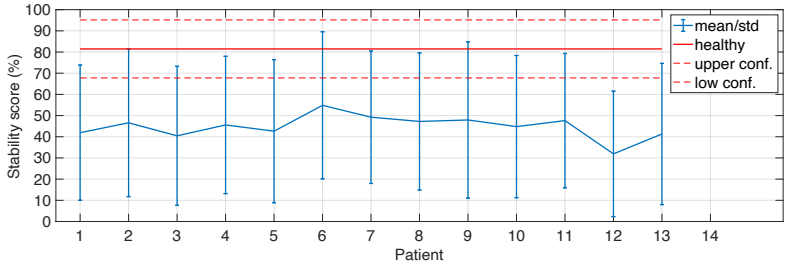

Figure 7: Average per patient stability performance w.r.t. the average stability of the healthy subjects and their confidence levels.

ted instability shall trigger a red alarm on the rollator's screen. In general, the mobility analysis results follow the respective POMA categorization of the patients highlighting the ability of this module to successfully assess measures essential for measuring rehabilitation progress.

4 Conclusions

This paper presents a multi-sensory multimodal framework, the i-Walk assessment system, that endows assistant platforms with the ability to successfully recognize human activities, understand audio-gestural intentions, monitor user's stability and mobility and assess rehabilitation progress giving meaningful feedback to the user. The i-Walk assessment system is extensively evaluated on a database of healthy subjects and patients where the presented results show quite high performance of the developed system components as well as the efficacy of the proposed framework not only to provide passive rollators with "intelligence", but also to be integrated into a general decision making strategy for natural and user-adaptive HRI of robotic assistant platforms.

References

1. DIAPLASIS Rehabilitation Center, https://www.diaplasis.eu/
2. Google cloud speech-to-text, https://cloud.google.com/speech-to-text/
3. RASA, http://rasa.com
4. RASA, http://https://github.com/RasaHQ
5. Robot Operating System (ROS), http://www.ros.org/about-ros/
6. Andriluka, M., Pishchulin, L., Gehler, P., Schiele, B.: 2d human pose estimation: New benchmark and state of the art analysis. In: CVPR (2014)
7. Bocklisch, T., Faulkner, J., Pawlowski, N., Nichol, A.: Rasa: Open source language understanding and dialogue management. In: NIPS (2017)
8. Cao, Z., Simon, T., Wei, S., Sheikh, Y.: Realtime multi-person 2d pose estimation using part affinity fields. In: CVPR (2017)
9. Chalvatzaki, G., Koutras, P., Hadfield, J., Papageorgiou, X.S., Tzafestas, C.S., Maragos, P.: Lstm-based network for human gait stability prediction in an intelligent robotic rollator. In: 2019 International Conference on Robotics and Automation (ICRA). pp. 4225–4232. IEEE (2019)
10. Chalvatzaki, G., Papageorgiou, X.S., Maragos, P., Tzafestas, C.S.: User-adaptive human-robot formation control for an intelligent robotic walker using augmented human state estimation and pathological gait characterization. In: 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 6016–6022. IEEE (2018)
11. Chalvatzaki, G., Papageorgiou, X.S., Maragos, P., Tzafestas, C.S.: Learn to adapt to human walking: A model-based reinforcement learning approach for a robotic assistant rollator. IEEE Robotics and Automation Letters **4**(4), 3774–3781 (2019)
12. Chalvatzaki, G., Papageorgiou, X.S., Tzafestas, C.S., Maragos, P.: Augmented human state estimation using interacting multiple model particle filters with probabilistic data association. IEEE Robotics and Automation Letters **3**(3), 1872–1879 (2018)
13. Chang, M., Mou, W., Liao, C., Fu, L.: Design and implementation of an active robotic walker for parkinson's patients. In: Proc. of SICE. pp. 2068–2073 (2012)
14. Chen, Y., Tian, Y., He, M.: Monocular human pose estimation: A survey of deep learning-based methods. Computer Vision and Image Understanding **192**, 102897 (2020)
15. Deriu, J., Rodrigo, A., Otegi, A., Echegoyen, G., Rosset, S., Agirre, E., Cieliebak, M.: Survey on evaluation methods for dialogue systems. Artificial Intelligence Review pp. 1–56 (2020)
16. Du, Y., Wang, W., Wang, L.: Hierarchical recurrent neural network for skeleton based action recognition. In: CVPR (2015)
17. Dubowsky, S., Genot, F., Godding, S., Kozono, H., Skwersky, A., Yu, H., Yu, L.S.: Pamm-a robotic aid to the elderly for mobility assistance and monitoring: a" helping-hand" for the elderly. In: ICRA (2000)
18. Efthymiou, N., Koutras, P., Filntisis, P.P., Potamianos, G., Maragos, P.: Multi-view fusion for action recognition in child-robot interaction. In: 2018 25th IEEE International Conference on Image Processing (ICIP). pp. 455–459. IEEE (2018)
19. Frizera-Neto, A., Ceres, R., Rocon, E., Pons, J.: Empowering and assisting natural human mobility: The simbiosis walker. Int. J. Adv. Robotic Syst. **8**(3) (2011)
20. Guler, A., Kardaris, N., Chandra, S., Pitsikalis, V., Werner, C., Hauer, K., Tzafestas, C., Maragos, P., Kokkinos, I.: Human joint angle estimation and gesture recognition for assistive robotic vision. In: ECCV (2016)
21. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural computation **9**(8), 1735–1780 (1997)
22. Jenkins, S., Draper, H.: Care, monitoring, and companionship: Views on care robots from older people and their carers. Int. J. Social Robotics **7**(5), 673–683 (2015)

23. Kulyukin, V., Kutiyanawala, A., LoPresti, E., Matthews, J., Simpson, R.: iwalker: Toward a rollator-mounted wayfinding system for the elderly. In: IEEE int. Conf. RFID (2008)
24. Leo, M., Furnari, A., Medioni, G.G., Trivedi, M., Farinella, G.M.: Deep learning for assistive computer vision. In: ECCV. pp. 0–0 (2018)
25. Lin, T., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.: Microsoft coco: Common objects in context. In: ECCV (2014)
26. Liu, J., Wang, G., Hu, P., Duan, L., Kot, A.: Global context-aware attention lstm networks for 3d action recognition. In: CVPR (2017)
27. Morris, A., Donamukkala, R., Kapuria, A., Steinfeld, A., Matthews, J.T., Dunbar-Jacob, J., Thrun, S.: A robotic walker that provides guidance. In: ICRA (2003)
28. Muro-De-La-Herran, A., Garcia-Zapirain, B., Mendez-Zorrilla, A.: Gait analysis methods: An overview of wearable and non-wearable systems, highlighting clinical applications. Sensors **14**(2), 3362–3394 (2014)
29. Papageorgiou, X.S., Chalvatzaki, G., Dometios, A.C., Tzafestas, C.S., Maragos, P.: Intelligent assistive robotic systems for the elderly: two real-life use cases. In: Proceedings of the 10th International Conference on PErvasive Technologies Related to Assistive Environments. pp. 360–365 (2017)
30. Papageorgiou, X.S., Tzafestas, C.S., Maragos, P., Pavlakos, G., Chalvatzaki, G., Moustris, G., Kokkinos, I., Peer, A., Stanczyk, B., Fotinea, E.S., et al.: Advances in intelligent mobility assistance robot integrating multimodal sensory processing. In: International conference on universal access in human-computer interaction. pp. 692–703. Springer (2014)
31. Paulo, J., Peixoto, P., Nunes, U.: Isr-aiwalker: Robotic walker for intuitive and safe mobility assistance and gait analysis. IEEE Trans. Hum. Mach. Syst. **47**(6), 1110–1122 (2017)
32. Perry, J.: "Gait Analysis: Normal and Pathological Function". Slack Incorporated (1992)
33. Piyathilaka, L., Kodagoda, S.: Human activity recognition for domestic robots. In: Field and Service Robotics. pp. 395–408 (2015)
34. Rezazadegan, F., Shirazi, S., Upcrofit, B., Milford, M.: Action recognition: From static datasets to moving robots. In: ICRA (2017)
35. Rodomagoulakis, I., Kardaris, N., Pitsikalis, V., Mavroudi, E., Katsamanis, A., Tsiami, A., Maragos, P.: Multimodal human action recognition in assistive human-robot interaction. In: ICASSP (2016)
36. Rodriguez-Losada, D., Matia, F., Jimenez, A., Galan, R., Lacey, G.: Implementing map based navigation in guido, the robotic smartwalker. In: ICRA (2005)
37. Roitberg, A., Perzylo, A., Somani, N., Giuliani, M., Rickert, M., Knoll, A.: Human activity recognition in the context of industrial human-robot interaction. In: Signal Inform. Proc. Assoc. Ann. Conf. (2014)
38. Shahroudy, A., Liu, J., Ng, T., Wang, G.: Ntu rgb+ d: A large scale dataset for 3d human activity analysis. In: CVPR (2016)
39. Sharkey, A., Sharkey, N.: Children, the elderly, and interactive robots. IEEE Robot. & Autom. Mag. **18**(1), 32–38 (2011)
40. Song, S., Lan, C., Xing, J., Zeng, W., Liu, J.: An end-to-end spatio-temporal attention model for human action recognition from skeleton data. In: AAAI Conf. Artif. Intell. (2017)
41. Stavropoulos, G., Giakoumis, D., Moustakas, K., Tzovaras, D.: Automatic action recognition for assistive robots to support mci patients at home. In: PETRA (2017)
42. Tinetti, M., Williams, T., Mayewski, R.: Fall risk index for elderly patients based on number of chronic disabilities. The American journal of medicine **80**(3), 429–434 (1986)
43. Tsiami, A., Filntisis, P.P., Efthymiou, N., Koutras, P., Potamianos, G., Maragos, P.: Far-field audio-visual scene perception of multi-party human-robot interaction for children and adults. In: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 6568–6572. IEEE (2018)

44. Tsiami, A., Koutras, P., Efthymiou, N., Filntisis, P.P., Potamianos, G., Maragos, P.: Multi3: Multi-sensory perception system for multi-modal child interaction with multiple robots. In: 2018 IEEE International Conference on Robotics and Automation (ICRA). pp. 1–8. IEEE (2018)
45. Veeriah, V., Zhuang, N., Qi, G.: Differential recurrent neural networks for action recognition. In: ICCV (2015)
46. Vemulapalli, R., Arrate, F., Chellappa, R.: Human action recognition by representing 3d skeletons as points in a lie group. In: CVPR (2014)
47. Vrigkas, M., Nikou, C., Kakadiaris, I.A.: A review of human activity recognition methods. Front. in Robotics and AI **2** (2015)
48. Wang, J., Liu, Z., Wu, Y., Yuan, J.: Learning actionlet ensemble for 3d human action recognition. PAMI **36**(5), 914–927 (2013)
49. Wang, P., Yuan, C., Hu, W., Li, B., Zhang, Y.: Graph based skeleton motion representation and similarity measurement for action recognition. In: ECCV (2016)
50. Zanfir, M., Leordeanu, M., Sminchisescu, C.: The moving pose: An efficient 3d kinematics descriptor for low-latency action recognition and detection. In: ICCV (2013)
51. Zhu, W., Lan, C., Xing, J., Zeng, W., Li, Y., Shen, L., Xie, X.: Co-occurrence feature learning for skeleton based action recognition using regularized deep lstm networks. In: AAAI Conf. Artif. Intell. (2016)
52. Zimmermann, C., Welschehold, T., Dornhege, C., Burgard, W., Brox, T.: 3d human pose estimation in rgbd images for robotic task learning. In: ICRA (2018)