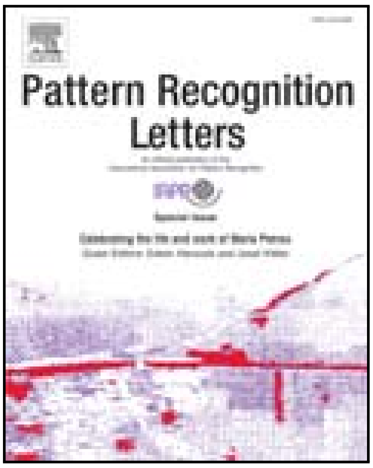
Accepted Manuscript



COMBAHO: A Deep Learning System for Integrating Brain Injury Patients in Society

Jose Garcia-Rodriguez, Francisco Gomez-Donoso, Sergiu Oprea, Alberto Garcia-Garcia, Miguel Cazorla, Sergio Orts-Escolano, Zuria Bauer, John Castro-Vargas, Felix Escalona, David Ivorra-Piqueres, Pablo Martinez-Gonzalez, Eugenio Aguirre, Miguel Garcia-Silviente, Marcelo Garcia-Perez, Jose M. Cañas, Francisco Martin-Rico, Jonathan Gines, Francisco Rivas-Montero

PII: S0167-8655(19)30053-4
DOI: https://doi.org/10.1016/j.patrec.2019.02.013
Reference: PATREC 7435

To appear in: *Pattern Recognition Letters*

Received date: 29 June 2018
Revised date: 28 January 2019
Accepted date: 10 February 2019

Please cite this article as: Jose Garcia-Rodriguez, Francisco Gomez-Donoso, Sergiu Oprea, Alberto Garcia-Garcia, Miguel Cazorla, Sergio Orts-Escolano, Zuria Bauer, John Castro-Vargas, Felix Escalona, David Ivorra-Piqueres, Pablo Martinez-Gonzalez, Eugenio Aguirre, Miguel Garcia-Silviente, Marcelo Garcia-Perez, Jose M. Cañas, Francisco Martin-Rico, Jonathan Gines, Francisco Rivas-Montero, COMBAHO: A Deep Learning System for Integrating Brain Injury Patients in Society, *Pattern Recognition Letters* (2019), doi: https://doi.org/10.1016/j.patrec.2019.02.013

Research Highlights (Required)

To create your highlights, please type the highlights against each `\item` command.

It should be short collection of bullet points that convey the core findings of the article. It should include 3 to 5 bullet points (maximum 85 characters, including spaces, per bullet point.)

- An interactive social robot for assistance and stimulation.
- Intelligent environment with abilities to monitor and learn actively.
- An outdoor assistant to help disoriented patients.
- Integration of several existing solutions and adaptation to the problem at hand.
- Deep learning solutions to face various challenges: memory loss, attention problems...

Pattern Recognition Letters
journal homepage: www.elsevier.com

COMBAHO: A Deep Learning System for Integrating Brain Injury Patients in Society

Jose Garcia-Rodriguez[a], Francisco Gomez-Donoso[a], Sergiu Oprea[a], Alberto Garcia-Garcia[a,**], Miguel Cazorla[a], Sergio Orts-Escolano[a], Zuria Bauer[a], John Castro-Vargas[a], Felix Escalona[a], David Ivorra-Piqueres[a], Pablo Martinez-Gonzalez[a], Eugenio Aguirre[b], Miguel Garcia-Silviente[b], Marcelo Garcia-Perez[b], Jose M. Cañas[c], Francisco Martin-Rico[c], Jonathan Gines[c], Francisco Rivas-Montero[c]

[a] *3D Perception Lab, University of Alicante, Spain*
[b] *University of Granada, Spain*
[c] *University Rey Juan Carlos, Spain*

ABSTRACT

In the last years, the care of dependent people, either by disease, accident, disability, or age, is one of the current priority research topics in developed countries. Moreover, such care is intended to be at patients home, in order to minimize the cost of therapies. Patients rehabilitation will be fulfilled when their integration in society is achieved, either in the family or in a work environment. To address this challenge, we propose the development and evaluation of an assistant for people with acquired brain injury or dependents. This assistant is twofold: in the patient's home is based on the design and use of an intelligent environment with abilities to monitor and active learning, combined with an autonomous social robot for interactive assistance and stimulation. On the other hand, it is complemented with an outdoor assistant, to help patients under disorientation or complex situations. This involves the integration of several existing technologies and provides solutions to a variety of technological challenges. Deep leaning-based techniques are proposed as core technology to solve these problems.

1. Introduction

Today's society faces a wide variety of social challenges. Among them, attention to the population in a situation of dependency stands out, a phenomenon that affects all ages, not just the elderly. This is demonstrated by the existence of people with congenital or acquired disabilities due to accidents, whether traffic, labor or domestic, and that form a significant number of dependents. The deterioration of the physical and cognitive abilities limits the autonomous life of the people, considering itself a primordial problem before which the developed countries have decided to act. In the specific case of the Spanish society, the demographic forecasts indicate that, in the year 2020, there will be about 1.5 million dependents in Spain, a figure that will increase due to the aging of the population. The studies carried out show that care for dependents must be carried out, especially in the family environment, in order to better integrate these people into daily life. In this sense, there is a growing interest in the analysis and study of how new technologies can help to improve the living conditions of dependent people. Actually, robotics is expanding its field of action, being implanted more and more at homes. This trend is known as social robotics and allows human-robot interaction following the rules, patterns and social behaviors, whose main objective is to assist, accompany and even facilitate rehabilitation tasks for dependents. There is a exponential growth in the development of systems of Ambient Assisted Living (AAL), which aim to prolong the time that people can live in a dignified way in their own home, increasing autonomy and self-confidence, improving their security and saving resources.

Social robotics and the AAL present a series of problems that today focus the work of different research groups. This reserch project is part of these new lines of research.

In the EU, the issue has been dealt with extensively in FP7 [1], **which is an European funding framework for researching and innovation**. Some of the systems that are already funded

[**] Corresponding author:
e-mail: `agarcia@dtic.ua.es` (Alberto Garcia-Garcia)

[1] https://ec.europa.eu/research/fp7

by these programs are: CompanionAble[2] whose objective is to take advantage of the synergy of robotics and Environmental Intelligence technologies and their semantic integration in the assistance environment of a caregiver. The same objective is pursued in the FLORENCIA projects[3], MOBISERV[4] and KSERA[5], that make use of an external network of sensors, placing the robot in a structured indoor environment. However, it is also generally ignored that, to be a true partner, robots have to be reliable, more intelligent and able to work in closer collaboration with humans[6]. The need to provide robots that can collaborate with people (companion robots) is explicitly included in the heart of Challenge 2: ”Cognitive Systems and Robotics of the EU Frame Programs”. For example, the Hobbit project[7] develops an assistance robot to help older people feel safe at home, whose goal is to keep people active with games and exercises, as well as detect emergency situations and act appropriately. The fundamental difference of this proposal with our system is that the robot is complemented with an intelligent monitoring system, and the range of people that can be treated is extended. On the other hand, the GiraffPlus project[8] proposes a complex system capable of monitoring the activity that takes place in a house through a network of sensors (distributed by the house or in the body of an elderly person), together with a robot that serves as an interface that allows different users, relatives, caregivers or health personnel to virtually visit the elderly person in their own home. Along the same lines, the Sphere project[9] in the United Kingdom has recently been funded with a large financial contribution.

In this project we focus on the care of patients with acquired brain injury ABI (Traumatism Cranioencephalic-TCE and Vascular Brain Accident-VBA) for being the second cause of affectation that generates dependence in developed countries.

In particular, we focus on the third phase based on increasing the autonomy and quality of life of the patient on his return home. In this phase patients suffer from the following problems:

1. Memory: frequent difficulty to remember tasks and to assimilate recent information.
2. Attention: attention problems are also a constant in people who have suffered an ABI, and are manifested as distractions and/or as a lack of ability to maintain attention.
3. Decision making: difficulty to initiate any task, doubting, make inadequate decisions or give ineffective responses facing certain situations.
4. Communication: difficulties in expression and comprehension. This is manifested in not answering the asked questions, in difficulties of expressing oneself, not understanding what is said, talking out of turn, abruptly changing the subject, etc.

In order to solve these problems we define a set of scientific challenges that must be addressed: action recognition, both coarse (movement of the patient) and fine grain (complex tasks such as brushing teeth) (1), 3D object recognition and manipulation (2), personal robotic assistance (3). Although there are already many works that attempt to solve these tasks, deep learning techniques have proved in terms of accuracy and robustness to be the most suitable for it. For example, the recognition of 3D objects with traditional methods has been studied, both 2D and 3D, but there are only a few works that apply robust learning techniques to the recognition of 3D objects. The results obtained in these first works suggests that there is still a long way to go in its development. Therefore, our baseline approaches to solve most of the presented challenges will be deep learning-based techniques.

Another example of tasks that we must address in this project is to develop techniques that are able to learn in continuous way (incremental learning) or learning without forgetting. Here unsupervised learning techniques are used to determine if the elements recognized in an unknown scene belong to previously known classes in the training set. In this case we will incrementally train the system with them, or assume new concepts for the network, which will be incorporated by modifying the architecture of the same automatically and transferring previously learned weights to allow classification of the new class, while maintaining the results in terms of accuracy of the previous network.

To sum up, our work aims to provide a rehabilitation and assistive system which features the following contributions:

- **An interactive social robot for assistance and stimulation.**
- **An intelligent environment able to actively monitor and learn activities.**
- **An outdoor assistant system integrated in the social robot to help disoriented patients.**
- **Deep learning-based solutions to face various rehabilitation challenges such as memory loss or attention problems.**

The rest of the paper is organized as follows: Section 2 reviews related works on deep learning techniques applied to the problems proposed in the project. Section 3 details the proposed system. Sections 4 presents a preliminary user study in a controlled environment. In Section 5 we present a performance evaluation of the system. Finally, Section 6 draws conclusions and directions for future work.

2. Related Works

In this section we will present state-of-the-art deep learning approaches for each task we will face.

[2] http://www.companionable.net
[3] http://www.florence-project.eu
[4] http://www.mobiserv.eu
[5] http://ksera.ieis.tue.nl
[6] http://cordis.europa.eu/ictresults
[7] http://hobbit.acin.tuwien.ac.at
[8] http://www.giraffplus.eu
[9] http://www.irc-sphere.ac.uk

2.1. Recognition of scenes

In the context of computer vision, the process of perceiving or recognizing a scene goes a step beyond the recognition of isolated objects. To achieve a complete visual understanding of a scene several complementary processes are required Li et al. (2009): classify the scene at a high level, recognize the different objects in it and their constituent elements and delimit them by segmentation. The recognition of scenes is one of the key capabilities required by the new wave of cognitive or social robots; A robotic system for this purpose must be able to interpret unstructured environment features of the real world.

2.2. Robust feature learning

One of the most pursued goals in this field has been to replace the manually designed descriptors, which require experts in the application domain, by multilayer networks capable of automatically learning those features. The solution to this problem was discovered by several groups independently in the 70s and 80s Rumelhart et al. (1988). This fact gave rise to a new branch of machine learning called Deep Learning LeCun et al. (2015). These are networks with multiple layers, different activation functions and connectivity that learn to map an input of a given size to an output also of pre-established size, usually a vector that contains the probability for each of the categories of the classification.

2.3. Recognition of volumetric scenes with convolutional neural networks

Motivated by the success of the application of convolutional neural networks (CNNs) for recognition tasks using RGB images, several research groups have extended this type of networks to use RGB-D data with depth information Lenz et al. (2015). This type of approach simply treats the depth channel as an additional channel along with the remaining RGB, so they are considered 2.5D methods. On the other hand, as evolution of these 2.5D methods, purely 3D architectures have also been developed to make convolutions on volumetric data Wu et al. (2015). However, none of the analyzed works makes use of purely volumetric information together with RGB data to take advantage of the virtues of both modalities. In addition, there is a need to provide the complete pose of the recognized object in a continuous solution space. Finally, these models have not been applied to recognize scenes under adverse conditions such as occlusions, disorder and changes in lighting.

2.4. Distributed and shared representations

The popularity obtained by deep learning in the field of computer vision is comparable to that achieved in the field of natural language processing (NLP), highlighting the use of recurrent neural networks (RNNs) capable of learning distributed representations Hinton (1986). These acquire an even greater importance when they are combined in other learning tasks such as image recognition using CNNs. Using this type of data from the NLP together with visual data such as images, deep networks can be used to learn ways to map both sources of data in a shared representation. At present, there are multiple works that combine both approaches to analyze a scene and answer questions about it Karpathy and Fei-Fei (2017), however none of these systems employ three-dimensional data for learning (useful for discerning the relative positions of objects and other geometric properties).

2.5. Unsupervised and online deep learning

One of the main problems or obstacles when applying the previously described methods, image analysis through CNNs, is the need for labeled datasets of considerable dimensions for the training of the aforementioned networks. Most of the information we have is not labeled, and for complex problems like the one we are dealing with, the manual annotation of volumetric data is an extremely tedious process. For this reason, the ability to learn from unlabeled data is one of the directions towards which a large number of research efforts are conducted Chen et al. (2016).

2.6. Action recognition

The recognition of human behavior in video sequences is an important research topic in the field of computer vision. Video surveillance, environmental intelligence, optimization of spaces, urban planning or life assisted by the environment are examples of applications in which an automated analysis of behaviors is increasingly needed. The different levels of understanding that can be found in the literature Moeslund et al. (2006) to analyze behavior are: movement, action, activity and behavior from lower to higher level of complexity. Despite this classification, many works treat activities and behaviors indistinctly. Different approaches to the problem have been proposed such as those reviewed in Turaga et al. (2008). However, many of the proposals focus on full recognition of human activities, but not on the prediction in terms of an early detection of what an individual will perform on the scene. The recognition of behavior can be seen as a classification problem. However, predicting the activity of a behavior involves inferring behavior using a subset of data from the entire activity The complex human activities or prediction of behavior have been extensively studied in recent decades, however it remains an open research topic. It is a more complex problem because the number of possibilities is higher compared to the prediction of a single action. The forecast period of the activity, presented in Kitani et al. (2012), carries out the prediction of the behavior through the semantic knowledge of the scene and the theory of optimal control. His experiments focus on the prediction of trajectories, but the proposal has been presented for general situations. Cao et al. present in Cao et al. (2013), the use of scattered coding and sub-samples of the sequence to predict subsequent activities of partially observed sequences.

2.7. Deep learning-based techniques for action recognition

In recent years, there has been a boom in methods based on deep learning architectures for the recognition of behaviors or actions in video sequences Baccouche et al. (2011). In particular, recurrent neural networks (RNN) Veeriah et al. (2015) and long and short-term memory networks (LSTM) Donahue et al. (2017), capable of processing sequential information efficiently using special connection schemes for the network architecture.

Such networks have the potential to model any temporal data series in which past states influence future ones, that is, they are able to memorize. It is for this reason that deep learning is postulated as a strong candidate for the analysis of behaviors in sequences.

2.8. Predictive learning

The ability to predict and therefore anticipate future events is an important and necessary attribute for intelligent decision making systems. From past experience and events, humans are able to predict effortlessly, for example, vehicle trajectories and pedestrian behavior. However, this remains an open challenge for today's computer vision and deep learning systems. In intelligent systems where real time is a key factor, the ability to anticipate can suppose a big improvement in terms of accuracy, reliability and robustness. The concept of predictive learning will determine the new frontier for intelligent systems, based on attributing the quality of ”common sense” to the machines themselves in order to improve real-time decision making.

In the literature, there are many works trying to predict future from low-level features such as pixels intensities of video frames Mathieu et al. (2015) to more high-level ones: semantic segmented video frames Neverova et al. (2017), trajectories, instance segmented frames Luc et al. (2018), scene dynamics Fouhey and Zitnick (2014), etc. We can also use future prediction to solve long-term planning problems Shalev-Shwartz et al. (2016).

In contrast of predicting future high-level scene properties, modeling raw RGB intensities is a very complicated task due to the high dimensionality of data. However, the prediction of low-level future is not always the key for decision-making systems, since for example, it is the semantic segmentation the cornerstone of those systems and represents one of the most complete forms of visual scene understanding. The prediction of abstract data is not only more accurate, but also it would be more useful for our tasks.

However,the main problem for training robust predictive models with high-level labeled data is the lack of large video datasets providing this information. There are two main ways to address this problem: training predictive systems in an unsupervised manner Misra et al. (2016) Pathak et al. (2017) or using large-scale synthetic datasets which are properly labeled. State-of-the-art unsupervised systems are still half way to being fully reliable and robust, and are limited to pretrain networks that will be netunned with labeled data. However, this introduces signicant improvements and it has been proven that these systems are, for example, able to learn scene dynamics Fouhey and Zitnick (2014).

3. COme BAck HOme Project

3.1. Project description

The care of dependent people, either by disease, accident, disability, or age, is one of the current priority research topics in developed countries. This care, besides of providing assistance and company, is considered even therapeutic. Moreover and in order to minimize the cost of therapies, it is intended to be at patients home. Patients rehabilitation will be fulfilled when their integration in society is achieved, either in the family or in a work environment. To address this challenge, the main scientific objective of this project is to promote health and well-being of society from the design, development and evaluation of an assistant for people with acquired brain injury or dependents to help them face the challenges of their illness raises in their full social integration. This assistant is twofold: at the patient's home is based on the design and use of an intelligent environment with capacities to monitor and active learning, combined with an autonomous social robot for interactive assistance and stimulation. On the other hand it is contemplated with an outdoor assistant, to help patients under disorientation or complex situations. This involves the integration of several existing technologies and provides solutions to a variety of technological challenges that these systems lead paired. In addition, we propose an experimental evaluation conducted by clinicians who assessed the effectiveness of the system to improve the quality of life of dependent people. Both, autonomy and the positive cognitive-affective state of the patient will be assessed.

3.2. Goals

To achieve the overall objective proposed, it is necessary to address certain scientific-technological challenges that are split into the following specific objectives: i) develop a robust intelligent monitoring system environment that allows locate and track, precisely, individuals in the scene; ii) develop a robust positioning and navigation system, to recognize and manipulate 3D objects of reduced size with a robot; iii) design a custom assistant to help the patient in situations of memory lapses, lack of orientation, motor difficulties, visibility reduction and other situations. This system will be trained to conduct and common scenarios indoors and outdoors; iv) to improve the assistant with natural interaction capabilities using new techniques of natural language processing combined with visual attention and deep learning. v) to perform a design of assistance and care rehabilitation stage, and identify metrics, pilot and final evaluation of the developed system with real patients and scenarios; finally vi) to publish and disseminate the results to the scientific community and related companies and associations in the area.

The expected results of the project are diverse. At scientific-technical level is expected to achieve significant advances in developed technologies. In terms of social impact, it is expected to improve the quality of life of patients. As for the economic impact, it is expected to obtain a functional system and its possible commercialization increasing technology transfer to society.

3.3. Project tasks

The tasks proposed to create the support systems are:

1. Ambient intelligence system: design and implement an intelligent environment based on low cost 3D visual sensors with the purpose of monitoring patients. This will allow precise tracking and localization in indoor environments, i.e. patient's home.

2. Robotics assistant system: design and develop a robust module for location and navigation through the scenario, and also, a module for recognize and manipulate small 3D objects. Both will be integrated aboard the robot.
3. Indoor patient assistant: design and develop a personalized assistant at patient's home based on deep learning techniques and specialized in temporal sequences to help the patient in situations of cognitive failures.
4. Human-robot interaction: implement an interaction module by which patients can give orders and communicate with the assistant robot by means of gestures. The gesture recognition module will be able to recognize successfully Schaeffer language.
5. Outdoor patient assistant: develop an outdoor patient assistance system using a wearable vision sensor and location information provided by GPS to facilitate patient's navigation.
6. Integration and testing: perform the integration of the different systems. For this purpose, we will design care scenarios in order to perform a final evaluation of the entire system with real patients.

3.4. Implementation

3.4.1. Ambient Intelligence System

The Ambient Intelligence System (AIS) is composed of several RGB-D sensors, small computers, and detection software. Sensors are installed close to the ceiling and cover the most interesting areas of the house. They are connected to small networked NUC computers with no screen as we can see in Figure 1.

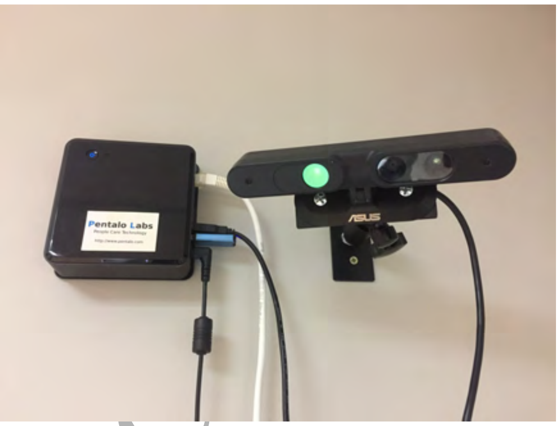


Fig. 1: Asus Xtion camera connected to a small networked NUC computer and installed near to the ceiling and in the corner of a room.

All sensors are calibrated (intrinsic and extrinsic parameters) and they work on the same spatial reference system. The idea of using depth sensors is to work 24/7, even at nights, in complete dark scenarios. This is a requirement for instance to detect people falling while trying to get to the toilet at night. Microsoft Kinect-1 and Asus Xtion sensors have been used in testing and experiments.

The detection software implements a distributed algorithm that has been designed to detect and track people in RGB-D images. It has been programmed in C++ and runs on Linux machines. Two different detection software algorithms have been developed: one based on background subtraction and a second one based on deep learning.

The first one includes several subsystems like background learning on each camera, background subtraction, clustering of foreground objects, and 3D tracking. Taking advantage of the depth sensor, it checks some spatial, volumetric and, aspect ratio constraints to discard false positives. A movement constraint was also added. It tracks "people candidates" in 3D space using a temporal filter similar to EKFs (Extended Kalman Filters). This filter was added to be robust to partial and temporal occlusions. As it tracks on the 3D space, instead of on each 2D image, the information coming from several sensors which cover the same area can be fused in a natural way. Each sensor provides person observations for the 3D tracker, which estimate new position of each person (the state) from those observations and the previous positions. In Figure 2 a typical person fall detection example is shown.

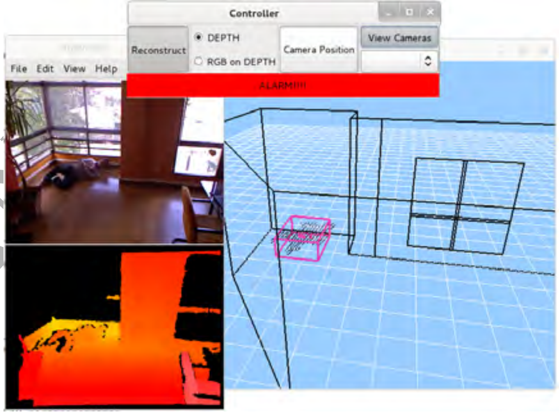


Fig. 2: Example of person falling detection.

The second algorithm runs a neural network to detect people in images and a tracking by detection module. The DarkNet framework has been used to train a fast YOLO network model for detection. In Figure 3 a typical deep learning-based person detection example is shown. The detector has been trained with the combination of PASCAL-VOC and a new dataset obtained from the logs of the detector based on background subtraction and 3D reasoning. This way the detector on the tested scenarios was refined and better performance was achieved.

Neural detections are accurate and robust, but usually are slow on low-end computers. In order to keep real time operation the tracking-by-detection is combined with a Lucas-Kanade feature tracking, which performs the people tracking in the frames between neurally processed images.

In addition, a new tool was created when developing this algorithm. It is named DetectionSuite [10] and allows training of neural networks using several datasets (like COCO, IMAGENET, PASCAL-VOC...). Several parsers have been developed. It also allows labeling and refining input data to generate

[10] http://jderobot.org/DetectionSuite

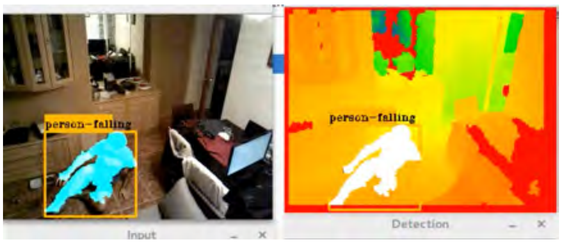


Fig. 3: Example of person falling detection with YOLO neural network

supervised tagged datasets. It also allows fair comparison of several network models using the same datasets and automatically computing performance measurements like IoU (Intersection Over Union).

The Ambient Intelligence System is connected to the other modules of the project through explicit ROS interfaces. On such interfaces the robot may ask, for instance, where in the house the person is located, and maybe go towards her.

3.4.2. Indoor Object Recognition

Object recognition is an indispensable feature for every robot, even more for an assistant one. In the context of this project, the robot must be capable of finding objects in the environment in order to retrieve it and bring it to the patient. Besides, it should also be capable of detecting which object is the patient manipulating, as this information is useful to predict which action is being performed. Action prediction and behavior analysis is another goal of the project, as detailed in Subsection 3.3. To fulfill this object recognition task, an ensemble of three different approaches is used as shown in Figure 4.

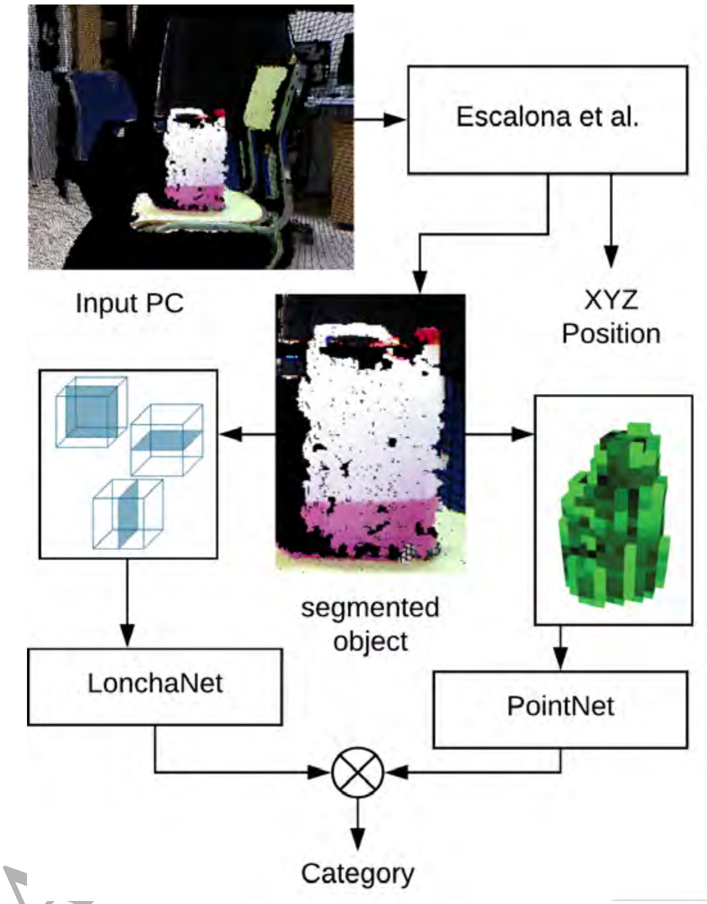


Fig. 4: Indoor object recognition pipeline consists of an ensemble of three different state-of-the-art approaches that provides high accuracy and robustness.

The first step of our object recognition pipeline is to infer the location of the object in the scene. In this regard, we implemented the approach described in Escalona et al. (2018). In this system, a robot is used to exhaustive inspect the environment by rotating over its Z axis. For each point cloud the robot has captured, color information is extracted and a CNN is used to infer what objects are present in the scene. Since there are various probabilities of finding different objects in the same frame, a probability profile is constructed for each object. In this probability profile, an increasing trend can be observed, which will reach a peak (since the object is appearing until it is fully observable) before gradually descend the probability of finding this object (as the object is fading away the scene that is captured by the robot). In this way, the system is able to isolate the 3D points of the object and to compute the three-dimensional position.

Then, a combination of two approaches is used to finally classify the object. First, PointNet Garcia-Garcia et al. (2016) is fed with the object's point cloud in order to classify it. This approach makes use of pure 3D convolutions to extract three-dimensional features that allow the system to classify the object. After that, LonchaNet Gomez-Donoso et al. (2017a) is used to classify the object once again. This approach extracts three slices from the point cloud, one slice per axis, and forwards them to the deep convolutional architecture. The outputs of the former systems are combined to produce the final classification of the object.

The first step is performed to localize the object in the 3D world. The next two architectures are used because PointNet provides high tolerance to occlusions, as explained in Garcia-Garcia et al. (2017), yet its classification accuracy is lower than the one provided by LonchaNet. By combining the output of the two architectures, we achieve high tolerance to occlusions while maintaining high classification accuracy.

3.4.3. Robot Navigation

Our navigation module is based on the navigation system offered by ROS Quigley et al. (2009) with several contributions. The ROS navigation system has become a standard in recent years, being used for robots equipped with laser sensors to navigate reliably and robustly through indoor environments. This system consists of three components (see Figure 5):

- **Map Server**. It is a component that offers a static occupancy map for the rest of the system. The map is stored as a grayscale image with a metadata file. Once loaded, the map is represented as a 2D occupancy grid, where its values in the range $[0-255]$ indicate the occupation (0 is free, 254 fully occupied and 255 unknown).

- **AMCL**. This component determines the position of the

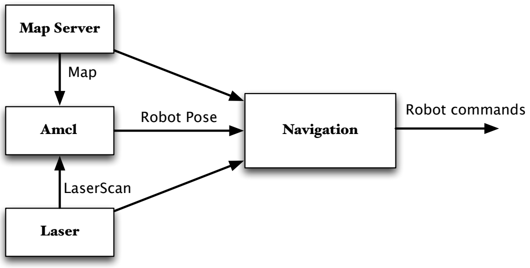


Fig. 5: ROS Navigation system.

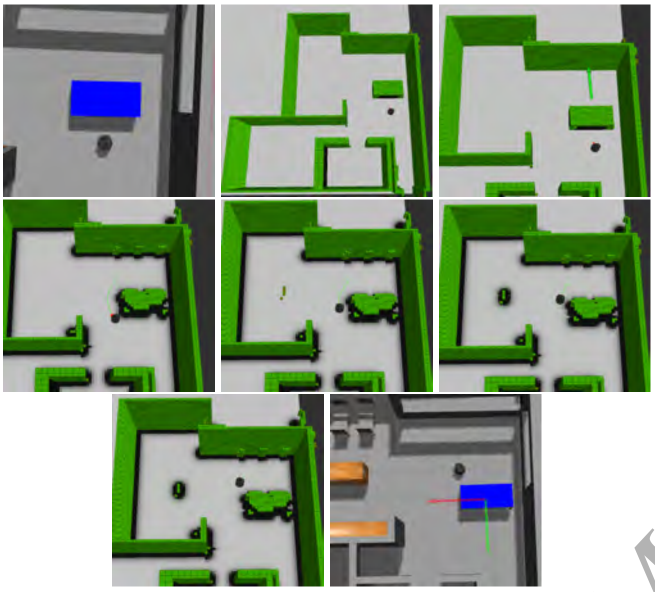


Fig. 6: Navigation module with 3D information and dynamics maps.

robot using its displacement, the readings of the laser, and the map obtained from the *map server*. It uses a version of Monte Carlo algorithm named KLD-Sampling Fox (2001).

- **Move Base**. This component implements navigation using both map information and robot position. It is organized in two levels: global and local navigation. Global navigation calculates the route from the position of the robot and its destination, taking into account aspects such as the robot's radius, security settings, and speed configuration. Local navigation tries to follow this route avoiding the possible dynamic obstacles that may be present.

The biggest problem with this system is that it only detects obstacles at the laser level, and this makes it not work correctly with obstacles below the laser line, or with tables and chairs, since it only detects the legs and can calculate navigation routes below them. In addition, this system does not modify the map once it has been created, in such a way that it does not take into account changes in the position of the furniture or closed doors.

For this reason we have improved this system in two ways: Use of 3D sensors and dynamic map update:

- **3D sensors**. Instead of using only the 2D laser, we also use the RGBD camera equipped in the robot. The 3D obstacles detected by the robot are stored in a structure called octomap Hornung et al. (2013), which stores space occupancy in voxels. This structure is stored by *map server*, which processes it and communicates it to the rest of the modules in the same way as the original component, but already showing as occupied the cells that would involve a collision with the robot at any height.

Fig. 7: Experiment with a static object in front of the robot.

- **Dynamic maps**. The *map server* starts from an static architectural map that only contains the elements that will never change: the walls. This component maintains a short-term map and a long-term map, besides of the static map. The detected obstacles are incorporated into the map in the short term, and if they persist over time, they are finally added to the map in the long term. Obstacles can also be eliminated when disappear. The *map server* offers as a map the combination of the three previous maps. In practice, this means that the robot incorporates into the architectural map the furniture of a house as it navigates, and that can be adapted to the changes that occur. The robot cal also avoid object which can be only detected using 3D information, as can be seen in Figure 6. Likewise, closed doors are incorporated into the map so that the robot can calculate new global routes.

- **Experiments** We have carried out several experiments to verify the correct functioning of the functionalities described above, fundamentally with the ability of the navigation system to incorporate new conditions in the environment that affect navigation into our dynamic map. We have carried several experiments to analyze how the navigation system incorporates static objects, ignore dynamic objects and how it adapts in the presence of closed doors:

  - In this experiment we want to check how static objects are incorporated into the map in which the navigation routes are calculated. In this experiment, the robot perceives a static object (Figure 7).

    At first, the object will be added to the short-term map and 17s later, when the value of the cells in short-term map reach 1, the long-term map will reflect this change.

    Figure 8 represents the evolution of the short-term map and long-term map when the robot perceive a new object. If the object is removed from the scene, the short-term map will reflects this change clearing the cells of the object and when the short-term map its clean, the cells of the object in long-term map will starts to decrease their value.

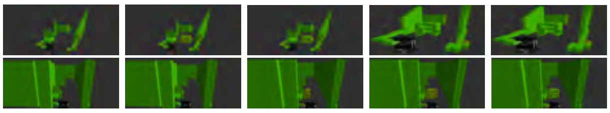


Fig. 8: Changes in cell's value in presence of static objects. Up, the short-term map and down the long-term map

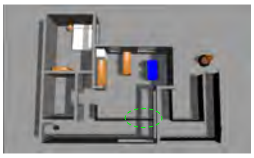


Fig. 9: Path blocked simulating a closed door.

- This experiment is designed to analyze how the navigation system adapts so that doors can be closed in the path of the robot, choosing an alternative path, if there is one. In this experiment, we ask the robot to go near to the blue table in the dinning room. It knows the map, so it can compute the path: go straight to the wall, turn left and go straight again to the table. In this experiment, we have blocked this way, simulating a closed door (Figure 9).

  The robot starts its way, but when it approximates to the closed door, map_{lt} starts to reflect this change. Path is recomputed, and the robot found an alternative way to reach its goal. The robot does not forget the closed door when the robot is looking away, because of this the robot will not use this door to calculate a new route.

The RoboCup@Home is an international competition that presents a common scenario where research groups from around the world compare their research in a common scenario. The scenario for this competition is a domestic environment where a dependent person lives. Several tests are proposed focused on developing the capabilities that a robot must have for a robot to be useful for this dependent person, among them, autonomous navigation.

We have participated in several editions of this competition. In particular, the navigation system described in this paper was used in Leipzig in 2016. A complete video can be found here[11].

Our mapping method showed several improvements over the techniques commonly used by most of the other teams. Many team needs to map the environment by guiding the robot along the arena while building the map. In this competition many teams participate, and most of the time the arena is crowded. It is complicated that it is completely empty to make the map with the robot. In addition, the referees periodically changed the position of the furniture. We measured the arena walls the first day to build an architectonic map, and periodically took a walk with the robot on the arena to update the furniture.

Fig. 10: Path followed.

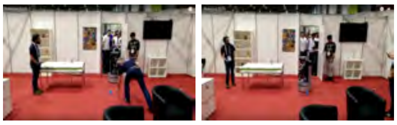


Fig. 11: Obstacle avoidance.

Fig. 12: Path recalculation when door is detected closed.

While navigation test, the referees closed the door (Figure 12). Our mapping system immediately incorporated the door as an obstacle, so the navigation system planned an alternate route to the goal position. This could not be done with the former mapping system.

3.4.4. Outdoor Scene Understanding

It is worth noting that the patient will not be assisted by the robot in outdoor environments but he/she will be helped by wearable technology. A full-HD color camera will be attached to the clothes and it will connect via Wi-Fi to the patient's smartphone, which will in turn use a 4G connection for data transference. As expected, the reduced computational power of current smartphones prevents in-situ deep learning uses, so the computation is performed in a remote server. The output of the deep models, which are executed in the remote server, will be forwarded back to the patient's smartphone in order to raise speech or haptic alerts (e.g., collision or architectonic barriers warnings).

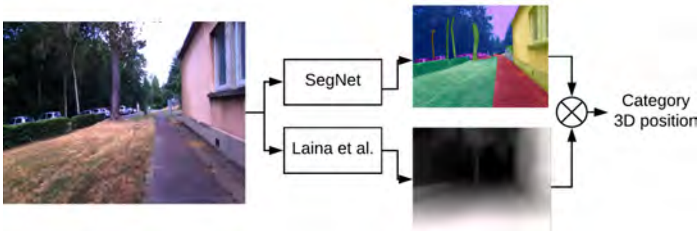


Fig. 13: Outoor scene understanding pipeline. Using a wearable camera, the patient is able to receive feedback about their surroundings.

In order to enable outdoor scene understanding and individual collision avoidance, we implemented the pipeline depicted

[11] https://www.youtube.com/watch?v=arUlQy7IT14

in Figure 13. First, a pixel-wise classifier performs inference on the wearable camera feed. This very same color frame and the previous one are both forwarded to a deep autoencoder that infers the correspondent depth map. The pixel-wise mask is used to extract potential collision subjects such as bikes, pedestrians, or poles. The predicted depth is used to compute the distance to the obstacle. As we already mentioned, two deep learning systems are jointly used: a semantic segmentation architecture and monocular depth estimation model.

First, in order to solve outdoor scene understanding, SegNet Badrinarayanan et al. (2017) is used. This deep learning architecture is able to provide pixel-level classification on color frames. We adopted the publicly available model as released by the authors, which is trained on the CamVid dataset Brostow et al. (2008) and additional public labeled data. This model is able to recognize the following categories: sky, building, pole, road, pavement, tree, sign symbol, fence, vehicle, pedestrian and bike. These categories are useful to detect potential obstacles present in outdoor environments. The global accuracy of this model is 86.8%, the class accuracy is 81.3% and the mean intersection over Union is 69.1% as reported by the authors.

However, the former system is only able to infer the position of the object in the image plane. In order to estimate their depth, some kind of three-dimensional information is needed.

As aforementioned, we adopted the approach proposed in Laina et al. (2016) to estimate the corresponding depth for each color image of the wearable camera. This architecture is built upon a ResNet50 He et al. (2015) scheme, but the last fully connected layer was replaced by a set of up-sampling blocks. These up-sampling layers are in charge of predict the corresponding depthmap for the input image. In this case, according to the authors, the architecture is intended to work on indoor environments, so the publicly available model is not suitable for our task. Nonetheless, we followed the paper details in order to train our own model from scratch with outdoor data.

The state-of-the-art outdoor-oriented datasets Ros et al. (2016); Menze and Geiger (2015); Schöps et al. (2017); Saxena et al. (2009) are either synthetic, reduced sized or yielded the point of view of a vehicle, rendering them virtually useless for out goal. So, we captured our own dataset which consist of 92616 RGB images and 46308 depthmaps distributed in 33 sequences that covers the outdoor surroundings of 38 different buildings of the University of Alicante campus. The ground truth was provided by a ZED Stereo camera. This device was attached to a head mount worn by a human in order to provide outdoor environments from a pedestrian point of view. Finally, we used this dataset to train the depth estimation architecture.

The training stage was performed under the following parameters: The model was initialized with the ImageNet pretrained Resnet50 weights. The learning rate was 0.01 and it gradually decreased every 6-8 epochs in a factor of 10. The batch size was 16. It was trained for 20 epochs. Data augmentation was performed as suggested in Eigen et al. (2014). The loss function was Reversed Huber Loss Zwald and Lambert-Lacroix (2012).

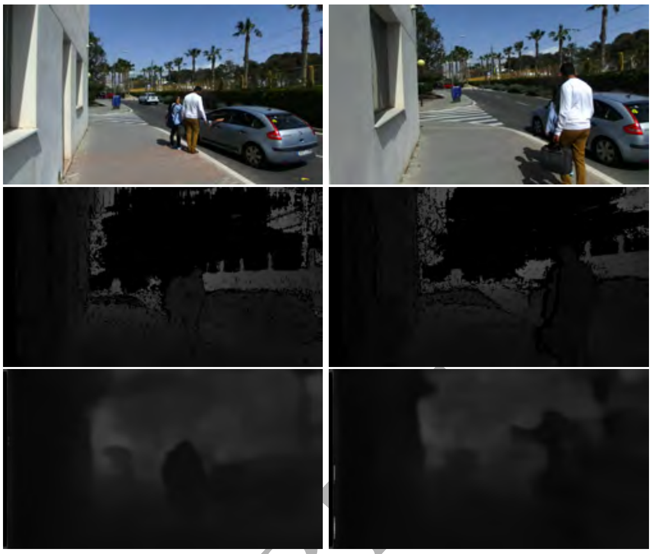


Fig. 14: Results of the depth estimation from monocular frames. First row shows two random samples, second row depicts the corresponding ground truth and third row represent the estimated depthmap as provided by the system. Images are shown for qualitative evaluation only.

Table 1 shows Mean Relative Error (MRelE) and the Root Mean Squared Error (RMSE) for the validation split and Figure 14 shows some results for qualitative evaluation purposes. It is worth noting that the predicted depthmaps are resized to fit the original input size in order to allow straightforward align with both color and pixel-wise classifications.

Architecture	MRelE	RMSE
OURS	0.173	2.124

Table 1: MRelE and RMSE errors obtained by the model trained with the UA-SOL dataset applied to the test split.

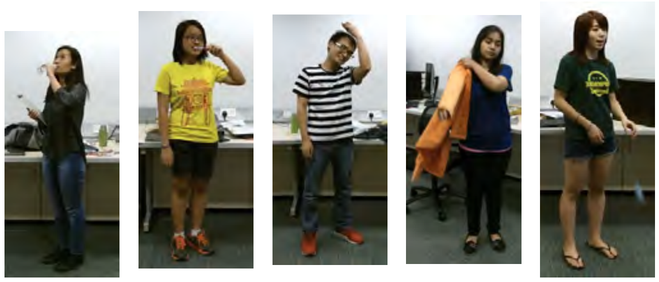
By using the predicted depth map, which is aligned with the color frame and the pixel-wise classifications, we can accurately obtain the distance of each pixel of the objects and raise collision alerts or provide guidance steps.

3.4.5. Human-Robot Interaction

The final system must be able to detect gestures from patients if they have serious verbal communication problems. For this purpose, a gesture detection module has been implemented for detecting Schaeffer language by means of which the patient can give orders to the assistant robot. The system is able to recognize 25 gestures with an accuracy of 93.13% using a three-layered LSTM network Sergiu et al., using as input, handcrafted features based on skeletal joint positions, angles and distances. One interesting feature is the early detection of a gesture. This is determined by the size of the sliding-window which should be large enough to discriminate between all the gestures. In literature, we found similar works about Schaeffer gesture recognition task such as Francisco et al.

3.4.6. Behavior Analysis

As we already mentioned, the recognition of human behavior in video sequences is an important research topic which has



(a) Drinking (b) Brushing teeth (c) Brushing hair (d) Wearing a jacket (e) Dropping something

Fig. 15: Actions to recognize

attracted the attention of many researchers Wu et al. (2017). This topic is a challenging task because of problems such as occlusions, cluttered backgrounds or viewpoints variations, so that there is not still a perfect system with a good performance in every situation.

Due to the amount of information contained in a video sequence we are going to deal with 2D images in this part of the project. Firstly, a single frame CNN approach has been implemented. This technique ignores temporal features and tries to classify each action with only a single frame. We have used an InceptionResnetV2 architecture pretrained on ImageNet. InceptionResnetV2 Szegedy et al. (2017) is one of the last architectures created by Google which is based on residual blocks. Our second implementation is based on 3D CNN Tran et al. (2015). This method consists of using a three-dimensional CNN which is an effective approach for spatio-temporal features. We have used the architecture of Tran et al. (2015) called C3D that it is composed by small $3 \times 3 \times 3$ convolution kernels and 3D pooling layers.

There exist several public datasets for action recognition such as KTH, WEIZMANN, MSR Action or NTU RGB+D Shahroudy et al. (2016). We have decided to use the last dataset due to it being the largest dataset with 60 classes and 56880 video samples acquired with a Kinect 2 sensor with 1920×1080 pixels resolution. Apart from the amount of data, this dataset has been collected with 40 different subjects and several viewpoints. In our project, we are interested in the analysis of the execution of certain activities by dependent or disabled persons. By analyzing how these persons are carrying out the activity, useful information can be obtained to help them. The first step is the detection of the activity in healthy people. At this time, five classes have been selected: drink, brushing teeth, brushing hair, wearing a jacket and drop an object. Figure 15 shows examples of theses actions.

Models have been trained with 3150 (630 per class) training videos and 795 validation videos. Once the model have been obtained it has been assessed using a test set with 795 videos which have been not previously seen by the system. Due to the high computational cost, the videos of the samples have been scaled down in resolution and frames per sample with respect to the original format. Final results on the test set are shown in Table 2.

Technique	Single frame CNN	3D CNN
Accuracy	0.534	0.862

Table 2: Different techniques results

The average accuracy of the single frame CNN approach is lower than the one from 3D CNN due to the fact that the former method is not considering temporal features of the video sequence while the latter method uses both spatial and temporal features of the samples. The value shown in Table 2 is the mean accuracy over the five classes so that it is interesting to obtain the confusion matrix to analyze the performance of this proposal. The confusion matrix for the five classes is shown in Tables 3 and 4 in absolute and relative values respectively.

Table 3: Confusion matrix in absolute values

Prediction/Actual	(a)	(b)	(c)	(d)	(e)
(a)	125	24	1	0	0
(b)	15	115	3	9	0
(c)	16	18	145	3	1
(d)	3	2	0	145	0
(e)	0	0	10	5	155

Table 4: Confusion matrix in relative values

Prediction/Actual	(a)	(b)	(c)	(d)	(e)
(a)	0.786	0.151	0.006	0	0
(b)	0.094	0.723	0.019	0.056	0
(c)	0.101	0.113	0.912	0.019	0.006
(d)	0.019	0.013	0	0.895	0
(e)	0	0	0.063	0.031	0.994

As it is shown by these tables, the obtained accuracy on every class is enough to the meet the requirements of the project. In order to improve the action recognition we plan on using Long Term Convolutional Networks (LRCN) Donahue et al. (2015). LRCN is a kind of architecture which combines a Convolutional Neural Network with a Recurrent Neural Network. This kind of recurrent network consists of complex recurrent modules and is capable of learning long-term dependencies so that we expect the accuracy to increase even if more classes are taken into account.

4. First User Study

The user study were develop in a controlled environment with the participation of patients as Primary Users (PUs) and relatives as Secondary Users (SUs). In this first experiment we tested the system with five patients in the last phase of the rehabilitation process and some relatives too. In the study we consider the evaluation of certain aspects such as: Usability, level of satisfaction, comfort, useful level, trusting level, privacy, and confident increase level in order to contribute to patient integration in society. Tests with questions related with the evaluation aspects were filled by users at the end of the study, which was carried out during

four weeks. In the trials, we evaluated the level of agreement with the questions and qualitatively measured it as: strongly disagree, disagree, neutral, agree, strongly agree and no opinion.

The questions included in the trial were:

1. Was the system easy to use?
2. Did the system acomplish your expectations?
3. Did you feel comfortable using the system
4. Was the the system useful?
5. Do you trust in the system reliability?
6. Did you feel your privacy respected?
7. Did the system contribute to your social integration?

Table 5 summarizes the users answers to questions 1-7 marking with stars the most popular responses among system users.

The conclusions that we extracted from the study and we plan to apply to our system to improve in further versions are:

1. Most of the patients found the system easy to use. The gesture interaction was the system that users found more difficult to use.
2. In general the users were positive related with their expectations about the system. The SUs were particularly impressed by the indoor and outdoor alert modules.
3. The users felt in general comfortable with the system. The system was tested in a controlled environment and not in patients homes which would change PUs and SUs opinions.
4. All the users found the system useful
5. In general the users trusted the system and find it respectful with their privacy.
6. Finally, all the users, PUs and SUs consider that the system facilitated and contributed to their social integration

Table 5: Users Study Questionaire Results

Quest. Answ.	Strong Dis.	Disagree	Neutral	Agree	Strong Agr.	N/A
1				⋆		
2				⋆		
3				⋆		
4					⋆	
5				⋆		
6				⋆		
7					⋆	

5. Performance evaluation

Our proposal uses different technologies in order to create a full assistant that covers all the necessities of an acquired brain injury patient. However, not all of them are integrated in a common framework. As explained before, the outputs of certain subsystems are used as inputs to feed other components. We are currently using ROS as our middleware of choice to integrate all subsystems.

Furthermore, almost all the described systems take advantage of deep learning techniques so they require considerable amounts of computational power. In order to improve the global performance of the system, we are currently modifying the system so that common models can be shared by different subsystems. The common models would be instantiated only once and would be used by different subsystems at the same time. By doing so, we can lower the overall computation requirements and reduce the computational cost.

Table 6: Table of the runtime for each single subsystem that composes our proposal. Note that some of the methods run simultaneously.

System	Processor	Runtime (ms)
Ambient Intelligence System		
BG Substraction	NUC	56.2
Person Detection (3D)	NUC	18.3
Person Detection (YOLO)	NUC	514
Communication Overhead	NUC	5.2
Accumulated Runtime(ms)		575.4
Indoor Object Recognition		
Escalona et al.	DLS1_GPU0	29.2
PointNet	DLS1_GPU0	24.6
LonchaNet	DLS1_GPU1	89.6
Communication Overhead	DLS1	2.1
Accumulated Runtime(ms)		120.9
Robot Navigation		
Localization and Mapping	Robot	58
Communication Overhead	Robot	1.7
Accumulated Runtime(ms)		59.7
Outdoor Scene Understanding		
SegNet	DLS1_GPU0	52.6
Laina et al.	DLS1_GPU1	55.2
Communication Overhead	DLS1	34.4
Accumulated Runtime(ms)		89.6
Human-Robot Interaction		
3L-LSTM	DLS1_GPU0	25.3
Communication Overhead	DLS1	2.7
Accumulated Runtime(ms)		28
Behavior Analysis		
3DCNN	DLS1_GPU1	98.3
Communication Overhead	DLS1	2.6
Accumulated Runtime(ms)		100.9

Nonetheless, the system is currently able to run in real time at interactive rates. As shown in Table 6, the indoor object recognition pipeline only takes 120.9 milliseconds, which means 8 FPS approximately. An inference step of the outdoor scene understanding module runs in 89.6 milliseconds (approx. 11 FPS). This frame rate is enough to provide real time information to the user.

As mentioned earlier, three main hardware setups are involved in our proposal. First, the NUC features an Intel i3-7100U running at 2.4GHz with 8GB DDR3 RAM. Then, a deep learning server (DLS1) is also used. This computer features an Intel i7-6800K running at 3.4GHz with 16GB DDR4 RAM and two main GPUs: Nvidia GTX1080Ti (DLS1_GPU0) and Nvidia Titan Xp(DLS1_GPU1). Finally, the robot yields its own on-board computer. It features an Intel Atom E3845 running at 1.91GHz with 4GB DDR RAM. The computers are interconnected by Gigabit Ethernet. The communication with the Outdoor Scene Understanding subsystem is performed over 4G LTE-Advanced.

6. Conclusion

In this paper, we presented some preliminary results of the COMBAHO project, which was designed to help integrating back into society disabled people with Acquired Brain Injury. The system is composed by ambient intelligence sensors and a social robot for indoor scenarios and a wearable camera and a smartwatch for outdoor situations. A number of deep learning-based solutions have been proposed and tested to face many challenges including: memory loss, attention problems, decision making, or communication among others. Since our current results were obtained in controlled environments, the main future research direction for the project involves testing the system in a wide variety of indoor and outdoor situations.

A first user study have been developed with some patients and relatives in a controlled environment with promising results. However, we are also working on a deployment test of the complete system and an exhaustive evaluation as well. We maintain a fluid communication with several brain injury associations and retirement homes for elders in which we successfully tested some of our systems independently and our former project, which is described in Gomez-Donoso et al. (2017b). Given the positive results, we are encouraged to perform new tests for our current approach.

Acknowledgements

This work has been funded by the Spanish Government TIN2016-76515-R grant for the COMBAHO project, supported with Feder funds. It has also been supported by the University of Alicante project GRE16-19 and the Valencian Government GV/2018/022. This work has also been supported by Spanish grants for PhD studies FPU15/04516 and ACIF/2017/243. Experiments were made possible by a generous hardware donation from NVIDIA.

References

Baccouche, M., Mamalet, F., Wolf, C., Garcia, C., Baskurt, A., 2011. Sequential deep learning for human action recognition, in: Proceedings of the Second International Conference on Human Behavior Unterstanding, Springer-Verlag, Berlin, Heidelberg. pp. 29–39.

Badrinarayanan, V., Kendall, A., Cipolla, R., 2017. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence .

Brostow, G.J., Shotton, J., Fauqueur, J., Cipolla, R., 2008. Segmentation and recognition using structure from motion point clouds, in: ECCV (1), pp. 44–57.

Cao, Y., Barrett, D., Barbu, A., Narayanaswamy, S., Yu, H., Michaux, A., Lin, Y., Dickinson, S., Siskind, J.M., Wang, S., 2013. Recognize human activities from partially observed videos, in: 2013 IEEE Conference on Computer Vision and Pattern Recognition, pp. 2658–2665.

Chen, G., Xu, R., Srihari, S.N., 2016. Sequential labeling with online deep learning: Exploring model initialization, in: Frasconi, P., Landwehr, N., Manco, G., Vreeken, J. (Eds.), Machine Learning and Knowledge Discovery in Databases, Springer International Publishing, Cham. pp. 772–788.

Donahue, J., Anne Hendricks, L., Guadarrama, S., Rohrbach, M., Venugopalan, S., Saenko, K., Darrell, T., 2015. Long-term recurrent convolutional networks for visual recognition and description, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 2625–2634.

Donahue, J., Hendricks, L.A., Rohrbach, M., Venugopalan, S., Guadarrama, S., Saenko, K., Darrell, T., 2017. Long-term recurrent convolutional networks for visual recognition and description. IEEE Trans. Pattern Anal. Mach. Intell. 39, 677–691.

Eigen, D., Puhrsch, C., Fergus, R., 2014. Depth map prediction from a single image using a multi-scale deep network, in: Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2, MIT Press, Cambridge, MA, USA. pp. 2366–2374.

Escalona, F., Gomez-Donoso, F., Cazorla, M., 2018. 3D Object Mapping Using a Labelling System. Springer International Publishing, Cham. pp. 579–590.

Fouhey, D.F., Zitnick, C.L., 2014. Predicting object dynamics in scenes, in: 2014 IEEE Conference on Computer Vision and Pattern Recognition, pp. 2027–2034.

Fox, D., 2001. Kld-sampling: Adaptive particle filters and mobile robot localization, in: In Advances in Neural Information Processing Systems (NIPS.

Francisco, G.D., Miguel, C., Alberto, G.G., Jose, G.R., . Automatic schaeffer's gestures recognition system. Expert Systems 33, 480–488.

Garcia-Garcia, A., Garcia-Rodriguez, J., Orts-Escolano, S., Oprea, S., Gomez-Donoso, F., Cazorla, M., 2017. A study of the effect of noise and occlusion on the accuracy of convolutional neural networks applied to 3d object recognition. Computer Vision and Image Understanding , –URL: http://www.sciencedirect.com/science/article/pii/S1077314217301182, doi:https://doi.org/10.1016/j.cviu.2017.06.006.

Garcia-Garcia, A., Gomez-Donoso, F., Garcia-Rodriguez, J., Orts-Escolano, S., Cazorla, M., Azorin-Lopez, J., 2016. Pointnet: A 3d convolutional neural network for real-time object class recognition, in: 2016 International Joint Conference on Neural Networks (IJCNN), pp. 1578–1584.

Gomez-Donoso, F., Garcia-Garcia, A., Garcia-Rodriguez, J., Orts-Escolano, S., Cazorla, M., 2017a. Lonchanet: A sliced-based cnn architecture for real-time 3d object recognition, in: 2017 International Joint Conference on Neural Networks (IJCNN), pp. 412–418. doi:10.1109/IJCNN.2017.7965883.

Gomez-Donoso, F., Orts-Escolano, S., Garcia-Garcia, A., Garcia-Rodriguez, J., Castro-Vargas, J.A., Ovidiu-Oprea, S., Cazorla, M., 2017b. A robotic platform for customized and interactive rehabilitation of persons with disabilities. Pattern Recognition Letters 99, 105 – 113. User Profiling and Behavior Adaptation for Human-Robot Interaction.

He, K., Zhang, X., Ren, S., Sun, J., 2015. Deep residual learning for image recognition. CoRR abs/1512.03385.

Hinton, G.E., 1986. Learning distributed representations of concepts, in: Proceedings of the Eighth Annual Conference of the Cognitive Science Society, Hillsdale, NJ: Erlbaum. pp. 1–12.

Hornung, A., Wurm, K.M., Bennewitz, M., Stachniss, C., Burgard, W., 2013. OctoMap: An efficient probabilistic 3D mapping framework based on octrees. Autonomous Robots Software available at http://octomap.github.com.

Karpathy, A., Fei-Fei, L., 2017. Deep visual-semantic alignments for generating image descriptions. IEEE Trans. Pattern Anal. Mach. Intell. 39, 664–676.

Kitani, K.M., Ziebart, B.D., Bagnell, J.A., Hebert, M., 2012. Activity forecasting, in: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (Eds.), Computer Vision – ECCV 2012, Springer Berlin Heidelberg, Berlin, Heidelberg. pp. 201–214.

Laina, I., Rupprecht, C., Belagiannis, V., Tombari, F., Navab, N., 2016. Deeper depth prediction with fully convolutional residual networks. CoRR abs/1606.00373.

LeCun, Y., Bengio, Y., Hinton, G.E., 2015. Deep learning. Nature 521, 436–444.

Lenz, I., Lee, H., Saxena, A., 2015. Deep learning for detecting robotic grasps. The International Journal of Robotics Research 34, 705–724.

Li, L.J., Socher, R., Li, F.F., 2009. Towards total scene understanding: Classification, annotation and segmentation in an automatic framework., in: CVPR, IEEE Computer Society. pp. 2036–2043. URL: http://dblp.uni-trier.de/db/conf/cvpr/cvpr2009.html#LiS09.

Luc, P., Couprie, C., LeCun, Y., Verbeek, J., 2018. Predicting future instance segmentations by forecasting convolutional features. CoRR abs/1803.11496.

Mathieu, M., Couprie, C., LeCun, Y., 2015. Deep multi-scale video prediction beyond mean square error. CoRR abs/1511.05440.

Menze, M., Geiger, A., 2015. Object scene flow for autonomous vehicles, in: Conference on Computer Vision and Pattern Recognition (CVPR).

Misra, I., Zitnick, C.L., Hebert, M., 2016. Shuffle and learn: Unsupervised learning using temporal order verification, in: Leibe, B., Matas, J., Sebe, N., Welling, M. (Eds.), Computer Vision – ECCV 2016, Springer International Publishing, Cham. pp. 527–544.

Moeslund, T.B., Hilton, A., Krüger, V., 2006. A survey of advances in vision-based human motion capture and analysis. Comput. Vis. Image Underst. 104, 90–126.

Neverova, N., Luc, P., Couprie, C., Verbeek, J.J., LeCun, Y., 2017. Predicting deeper into the future of semantic segmentation. CoRR abs/1703.07684.

Pathak, D., Girshick, R., Dollár, P., Darrell, T., Hariharan, B., 2017. Learning features by watching objects move, in: CVPR.

Quigley, M., Conley, K., Gerkey, B.P., Faust, J., Foote, T., Leibs, J., Wheeler, R., Ng, A.Y., 2009. Ros: an open-source robot operating system, in: ICRA Workshop on Open Source Software.

Ros, G., Sellart, L., Materzynska, J., Vazquez, D., Lopez, A., 2016. The SYNTHIA Dataset: A large collection of synthetic images for semantic segmentation of urban scenes.

Rumelhart, D.E., Hinton, G.E., Williams, R.J., 1988. Neurocomputing: Foundations of research, MIT Press, Cambridge, MA, USA. chapter Learning Representations by Back-propagating Errors, pp. 696–699.

Saxena, A., Sun, M., Ng, A.Y., 2009. Make3d: Learning 3d scene structure from a single still image. IEEE Trans. Pattern Anal. Mach. Intell. 31, 824–840.

Schöps, T., Schönberger, J.L., Galliani, S., Sattler, T., Schindler, K., Pollefeys, M., Geiger, A., 2017. A multi-view stereo benchmark with high-resolution images and multi-camera videos, in: Conference on Computer Vision and Pattern Recognition (CVPR).

Sergiu, O., A., G.G., S., O.E., V., V.M., A., C.V.J., . A long short-term memory based schaeffer gesture recognition system. Expert Systems 35, e12247. E12247 10.1111/exsy.12247.

Shahroudy, A., Liu, J., Ng, T.T., Wang, G., 2016. Ntu rgb+ d: A large scale dataset for 3d human activity analysis, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1010–1019.

Shalev-Shwartz, S., Ben-Zrihem, N., Cohen, A., Shashua, A., 2016. Long-term planning by short-term prediction. CoRR abs/1602.01580.

Szegedy, C., Ioffe, S., Vanhoucke, V., Alemi, A.A., 2017. Inception-v4, inception-resnet and the impact of residual connections on learning., in: AAAI, pp. 4278–4284.

Tran, D., Bourdev, L., Fergus, R., Torresani, L., Paluri, M., 2015. Learning spatiotemporal features with 3d convolutional networks, in: Proceedings of the IEEE international conference on computer vision, pp. 4489–4497.

Turaga, P.K., Chellappa, R., Subrahmanian, V.S., Udrea, O., 2008. Machine recognition of human activities: A survey. IEEE Trans. Circuits Syst. Video Techn. 18, 1473–1488.

Veeriah, V., Zhuang, N., Qi, G., 2015. Differential recurrent neural networks for action recognition. CoRR abs/1504.06678.

Wu, D., Sharma, N., Blumenstein, M., 2017. Recent advances in video-based human action recognition using deep learning: A review, in: Neural Networks (IJCNN), 2017 International Joint Conference on, IEEE. pp. 2865–2872.

Wu, Z., Song, S., Khosla, A., Yu, F., Zhang, L., Tang, X., Xiao, J., 2015. 3d shapenets: A deep representation for volumetric shapes., in: CVPR, IEEE Computer Society. pp. 1912–1920.

Zwald, L., Lambert-Lacroix, S., 2012. The berhu penalty and the grouped effect .