Combined Vision and Wearable System for Daily Activity Recognition

LOIZZO, FGC, FIORINI, L, SORRENTINO, A, DI NUOVO, Alessandro <http://orcid.org/0000-0003-2677-2650>, ROVINI, E and CAVALLO, F

Combined Vision And Wearable System For Daily Activity Recognition

Federica G. C. Loizzo[1,2], Laura Fiorini[3,5], Alessandra Sorrentino[1,2], Alessandro Di Nuovo[4], Erika Rovini[1,2], and Filippo Cavallo[3,1,2]

[1] The BioRobotics Institute, Scuola Superiore Sant'Anna, Pontedera (Pisa), Italy
[2] Department of Excellence in Robotics and AI, Pisa, Italy
[3] Department of Industrial Engineering, University of Florence, Florence, Italy
[4] Sheffield Hallam University, Sheffield, United Kingdom
[5] laura.fiorini@unifi.it

Abstract. Social assistive robotics aims at improving the quality of life of elderly people and caregivers. Human Activity Recognition (HAR) is one of the capabilities the assistive robot should be endowed with, to allow aged people to independently live in their homes. This work deals with the problem of performing HAR by employing two wearable inertial sensors and one RGB-D camera, mounted on the social robot Pepper. Specifically, the main purpose is to prove that Pepper robot is able to correctly recognize daily living activities by exploiting the information coming from the RGB-D camera and one inertial sensor placed on the index finger of the subject. Ten users were asked to perform ten activities while wearing an inertial glove, SensHand, and while being recorded by the camera. Two different perspectives of the robot were studied to understand if a good activity recognition could be obtained when the robot is in front of the person and on his side. The results show that almost the same recognition performances are obtained when combining the visual sensor, no matter the chosen perspective, with the inertial sensor only on the index (95%), with respect to the fusion of the same camera with the inertial sensor on the index and on the wrist (96%). This supports the conclusion that elderly people could just wear a small ring on the index finger to allow the robot to recognize their activities, taking advantage from a system which is comfortable and easy-to-wear.

Keywords: HAR · Gesture Recognition · Inertial Sensors · Visual Sensors · Machine Learning.

1 Introduction

Over the past decade, major neurocognitive disorder (NCD) has become a public health priority. The increasing number of people living with major NCD by 2050, according to the World Health Organization, raises the question of care, together with the risk of hospitalization, nursing home placement and the burden of professional and informal caregivers. As a consequence, it leads to increased care

costs. New initiatives based on psycho-social interventions, such as social-robot-based therapy, have been proposed as alternative solution to improve the quality of life of patients and caregivers [16,9,4]. Human Action Recognition (HAR) plays a vital role in the field of Human-Robot Interaction and it is widely researched for its potential applications [3]. It refers to an area of research that mainly involves automatic detection, recognition, and analysis of human actions from the data obtained from different types of sensors. Based on the specific applications, there are several activities that the HAR system is able to recognize. Activities of Daily Living (ADL) have given great importance to the monitoring of elderly people. In particular, in an Ambient Assisted Living (AAL) monitoring people while performing normal daily activities is essential [18]. In a AAL, activities like eating or drinking can be very important to help people keeping a healthy lifestyle, facilitating them to live longer in their family residential environments [11]. Different sensor modalities are employed in the HAR field. These include mainly RGB-D cameras and inertial wearable sensors. RGB-D video cameras are widely available and cost effective. They provide rich texture information of the scene and they are easy to operate. However, the vision-based approach is challenging many issues such as background clutter, occlusion, camera position, subject variations in performing actions and they are limited to a constrained space defined by the camera position and settings. To address such challenges, wearable inertial sensors are introduced to perform human action recognition. These include accelerometers and gyroscopes. This sensor technology has enabled coping with a much wider field of view as well as changing lighting conditions. Thanks to the lowering in the energy consumption and the increasing in the computational power of inertial sensors, long-term recordings have been enabled. These sensors allow to receive information directly from the movement of the users, detecting also fast and subtle movements without forcing them to stay in front of a camera. However, wearable inertial sensors have limitations as well. One of the main limitations is the sensor drift that may occur during long operation times; moreover, measurements are sensitive to sensor location on the body. In addition, for human action recognition, they require to be worn by subjects performing the actions, which creates the disadvantage of intrusiveness or inconvenience for the subjects. Even if a typical human action recognition system uses a single sensor, no single sensor modality can cope with various situations that may occur in real scenarios. One way to improve the performance of the human action recognition systems is to combine data from these two different modality sensors considering that images from a visual sensor and inertial signals from a wearable sensor provide complementary information. For example, images capture global (or full body) movement attributes while inertial signals capture fine movements, leading to a more robust recognition [6].

Therefore, the aim of this work is to combine inertial and visual data to obtain a system which can offer a robust activity recognition. Specifically, the main goal of this study is to go further the state of the art by evaluating whether an inertial sensor placed on the index finger, combined with visual data, is good enough to perform recognition of daily living activities. Skeleton data were obtained from

a RGB-D camera mounted over a social humanoid robot, Pepper, and they were combined with the inertial data acquired by a wearable glove, SensHand.

In a real-case scenario, a social robot could be very important to monitor the status of older people at home. Indeed, even if gesture recognition can be achieved by only exploiting fixed cameras and inertial data, it is much better that the camera, ideally characterized by robot's perspective, can move in the environment following the elderly person when required; indeed, it is not feasible to think of mounting several cameras to cover all the possible perspective in the environments. In this sense, a not-fixed camera mounted on the robot could overcome this limitation and could always adapt its point of view changing its perspective when required (i.e. when the older person is moving and/or changing activity or room). It is also very important to highlight that the interaction of an elderly person with a social robot allows the former to have company and not feel alone [19] [7].

In the proposed work, it is intended to simultaneously evaluate the performances of the system in two real-case scenarios, i.e. when the robot is in front of the person and when it is on the side. For this reason, during the experimental phase, two cameras were mounted frontally and laterally with respect to the subject performing the activity, to simulate the sight of the robot from two different perspectives. Such a system would improve the recognition rate of daily living gestures, by allowing the caregivers to monitor elderly people, and in particular people living with neurocognitive disorders, in any scenario.

The structure of the paper is herein presented: in Section 2 a general overview of the related works is provided, while in Section 3 the architecture of the system and the approach followed in this work are explained. Finally, in Section 5 and Section 6 the results of the previously performed data analysis are respectively presented and discussed.

2 Related Works

Several works focus on daily living activity recognition based on performed gestures detected by inertial and visual sensors. In their work, Dawar et al. [8] employed a wearable inertial sensor on the wrist and a Kinect v2 camera to recognize smart TV gestures. Acceleration and rotation signals from inertial sensors and skeleton data from depth cameras were extracted to train a Variable length Maximum Entropy Markov Model classifier. Action detection and recognition were performed continuously in real-time with the aim of separating actions of interest from actions of non-interest. Different scenarios were compared (subject-specific vs. subject-generic) with different values of the threshold probability p. The best performance in both the scenarios was observed at the threshold probability of p=0.45, obtaining 92% of precision. Wearable depth cameras for on-body activity recognition in home environment were used by Voigt et al. [21] to recognize 10 daily living activities. In particular, the Google Project Tango platform, which provides both a depth and an inertial sensor, was employed. After segmenting the signal, for each segment mean and standard deviation were

extracted as basic temporal features for all the sensors. With a total of nine features, they were able to achieve accuracy levels $> 90\%$. However, due to the size and weight of the platform, the system is not comfortable enough to be worn over long periods of time. Many works have combined inertial and depth sensors to other devices to improve gesture recognition. Li et al. [12] considered tri-axial accelerometers, micro-doppler radar and Kinect depth cameras to classify 10 different activities. Fusing information from the three sensors the classification accuracy reached 86.9% with the quadratic-kernel SVM classifier, and up to 91.3% using an ensemble classifier. Manzi et al. [15] aimed to recognize ten daily living activities using data from inertial sensors, worn on the index finger and on the wrist, from a depth camera mounted on a mobile robot and from the robot position, since the platform was able to self-localize in the environment. The classifier used three different types of features: user location, provided by the navigation module of the mobile platform, skeleton activity features, extracted from the raw skeleton data, and inertial features: mean, standard deviation, variance, mean absolute deviation, root mean square, energy, and IAV (integral of the magnitude of the acceleration vector), extracted from the accelerations' signals. Different combinations were tested and a decision-level fusion was applied. In their best configuration, namely fusion of depth camera, IMUs on wrist and index finger and location, accuracy levels of 70% were achieved. Supervised and unsupervised techniques are both used for classification. Usually a supervised classifier is used when the number of label and the actions to be recognized are already known. In particular, the most common ones are Support Vector Machine (SVM) [10], [12], the Random Forest (RF) [15], [10], [12], [21] and the K-Nearest Neighbors (KNN) [21]. Unsupervised approaches (e.g. k-means, Self-Organizing Map, and Hierarchical Clustering) were compared with supervised ones (RF, Multilayer Perceptron (MLP) and SVM) in [18]. The results reported in [18] about the intra-subject analysis, obtained as the mean value of 12 subject-dataset, were comparable with the results of the supervised analysis conducted with the 10-fold cross validation approach.

3 System Architecture

One of the aims of this work is to develop a not intrusive technology that can be adopted in real AAL scenarios as solution for monitoring the activity of elderly people. The activity recognition relies on a multimodal system composed by a wearable glove (i.e. Senshand) and a social humanoid robot (i.e. Pepper), as shown in Fig. 1.

SensHand is composed of four inertial modules positioned in correspondence to the wrist and to the thumb, index, and middle finger. Each module is composed of a complete 9-axis inertial sensor (6-axis geomagnetic module LSM303DLHC and 3-axis digital gyroscope L3G4200D, STMicroelectronics, Italy) and includes a microcontroller (ARM®-based 32-bit STM32F10RE MCU, STMicroelectronics, Italy) which can acquire, filter and store data at a frequency of 100 Hz [20]. Each module is able to measure metrics and parameters related

to posture, orientation and movement of the human hand. It is very easy to wear and to use thanks to its miniaturised and light structure; it is independent from the physical build of the person wearing it and from artifacts caused by the movement, making it suitable for remote rehabilitation and self-monitoring [20]. The entire system weights about 50 grams and its dimensions correspond to 3x4 cm as regards the wrist module and 1.5x1.5 cm for those on the fingers.

The system integrates a Pepper robot, which is the world's first social humanoid robot able to recognize faces and basic human emotions [2]. It is characterized by a multimodal sensing (i.e. touch sensors, infrared, cameras and sonars) thanks to which it can interact with people and move in an autonomous way. To enrich the visual capability of the robot, a RGB-D camera (i.e. Intel Realsense) is mounted on its chest over its tablet.

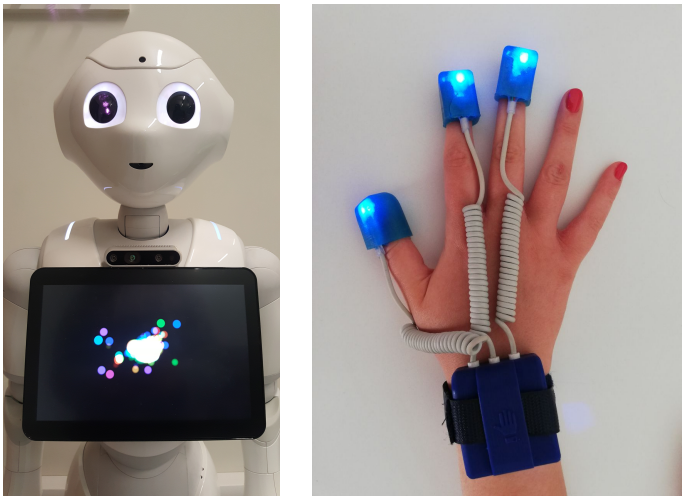

Fig. 1: Instrumentation employed: on the left, Pepper with the depth camera mounted above the tablet, and on the right, SensHand wearable glove.

The connection to the devices has been established via Bluetooth to the SensHand and via WiFi to the robot. Inertial data from the wearable glove and skeleton data from the cameras have then been integrated through a Python interface; in particular, two Python executables have been created to start data acquisition and transmission from the glove, while visual data have been acquired using the Robot Operating System (ROS) framework.

3.1 Experimental Protocol

The choice of the activities was based on the comparison between two public datasets, the Cornell Activity Dataset (CAD-60 and CAD-120) and the MSR

Daily Activity 3D Dataset. As result, ten activities in common between the two datasets were chosen, in such a way that they were similar in pairs (see Tab. 1).

Table 1: Description of gestures performed for the experimental session.

Activity	**Description**	**Position**
EF: Eat with the fork	Take the fork from the table, eat and put the fork back continuously	Sitting on the chair
DG: Drink from a glass	Take a glass from the table, drink and put it back repeatedly	Sitting on the chair
BT: Brush teeth	Take the toothbrush, brush teeth and put it back	Sitting on the chair
UL: Use laptop	Type on the keyboard with both hands	Sitting on the chair
WP: Write on a paper	Take a pen and write on a paper continuously	Sitting on the chair
TP: Talk on the phone	Take the phone, talk on it and put it back	Sitting on the chair
WK: Walk	Walk forward and backward repeatedly	Standing
SB: Sweep with the broom	Take the broom, sweep and put it back at the end	Standing
RC: Relax on the couch	Sit comfortably on the couch and relax	Sitting on the couch
RB: Read a book	Take the book, read it and turn pages repeatedly	Sitting on the couch

The experimental protocol consisted in the enrollment of 10 healthy participants, half males and half females, right-handed, from 19 to 44 years old. The experimental phase of this work was conducted in Sheffield (England), in the Smart Interactive Technology (SIT) research laboratory of Sheffield Hallam University. Study, design, and protocol, including subject privacy and sensitive data treatment, were approved by the Ethics Committee of Sheffield Hallam University. At the beginning of the experimental session, written informed consent was obtained from the participants. As a token of gratitude, participants received an Amazon e-voucher of £10 after successfully completing the experiment. During the experimentation, each subject simulated the ten activities, each for one minute, by wearing one SensHand glove on the dominant hand. The session was recorded by two cameras, one mounted over the robot and one located on the lateral side of the participant (Fig. 2) to acquire data from two different points of view, saving time, instead of asking the users to perform twice the protocol. The lateral camera is the same as the one mounted on Pepper and it was placed at the same height from the ground. The Pepper robot gave instructions about the action to perform and how to perform it. It is worth mentioning that the participants were left free to grab the objects and act in the way they preferred, so

no instruction was given in that sense. During the acquisition, each activity was labeled manually by an operator using an ad-hoc web interface. In particular, the interface has been appropriately created through an HTML code. It allowed to connect the sensors and, once selected the activity, to start the simultaneous data acquisition from the glove and the cameras.

At the end of the experimental trial, the participants were asked to fill in the System Usability Scale (SUS) questionnaire to assess the system usability. It consists of ten items with a five-point attitude Likert scale, providing a global view of subjective assessments of usability. A value equal or higher than 68 means that a certain technology is usable [1].

Fig. 2: Experimental setup in the Smart Interactive Technology (SIT) research laboratory of Sheffield Hallam University.

4 Data Analysis

The proposed activity recognition is performed on several steps. Firstly, data from the glove and the cameras have been analyzed on their own. This phase involved the extraction of the features from the sensors: the skeleton coordinates and inertial features. Then, the extracted features were organized in a database. The activity recognition was performed by employing supervised machine learning algorithms on unimodal data, collected in the dataset, and multimodal data, obtained by combining the previous ones. The multimodal classification was implemented at fusion-at-feature-level [22].

4.1 Pre-processing And Feature Extraction For IMU

Since the main frequencies of the inertial signal were between 0 and 5 Hz, a 4th order digital low-pass Butterworth filter was used, setting the cut-off frequency

at 5 Hz, similarly to [17]. In particular, acceleration and angular velocity data were filtered on their single components (x, y, z) and they were concatenated computing the Euclidean norm. According to the results obtained in [17], only the data coming from the wrist and index finger sensors of the glove have been used for the activity recognition.

Inertial data were segmented by 50 %-overlapping moving windows with a size of 3 seconds, considering that some individual actions were very short. For each window, many features were extracted. Tab. 2 shows the features extracted from acceleration and angular velocity signals. The final dataset has been composed by 10 features inherent to acceleration values and 6 features to angular velocities, for both wrist and index finger, for a total of 32 features.

Table 2: Features extracted from inertial data in time (t) and frequency domain (f).

Data	**Extracted Features**	
Acceleration	Mean value (t)	Skewness (t)
	Standard Deviation (t)	Kurtosis (t)
	Variance (t)	SMA: Signal Magnitude Area (t)
	MAD: Mean Absolute Deviation (t)	Normalized Jerk (t)
	RMS: Root Mean Square (t)	Power (f)
Angular Velocity	Mean value (t)	Mean Absolute Deviation (t)
	Standard Deviation (t)	Root Mean Square (t)
	Variance (t)	Power (f)

4.2 Pre-processing And Feature Extraction For cameras

As concerns RGB images analysis, the Openpose software [5] was employed to obtain the skeleton features. In particular, 25 keypoints were estimated for the body (see Fig. 3), where each of them represents the (x, y) pixels' coordinates of the joints. Some preliminary results show that a reduced set of joints could improve classification performances [13]. In the present study, a restricted set of joints was selected, namely composed by: head, neck, hands, feet and torso, which has been shown to be the most discriminative for activity recognition [14].

A normalization step was applied to the extracted features: the original reference frame was moved from the camera to the torso joint, and the joints were scaled with respect to the distance between the neck and the torso joint [14,15]. This normalization procedure yields data which are independent with respect to the person's specific size and to the relative position of the camera. Formally, considering a skeleton with N joints, the skeleton feature vector, f, is defined for each frame as in Eq. (1):

$$f = [j_1, j_2, \ldots, j_i, \ldots, j_{N-1}], \tag{1}$$

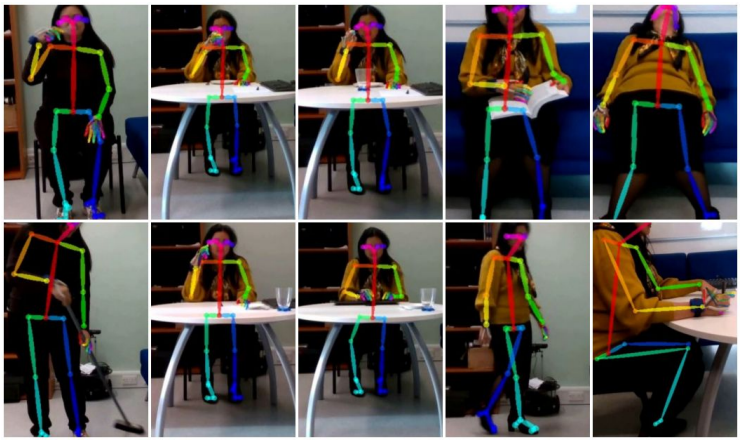

Fig. 3: Extracted skeleton for the 10 activities.

where each j_i contains the normalized coordinates of the i^{th} joint J_i detected by the sensor. Finally, considering all the frames, j_i expands to a vector and it is defined as in Eq. (2):

$$\mathbf{j}_i = \frac{\mathbf{J}_i - \mathbf{J}_0}{||\mathbf{J}_1 - \mathbf{J}_0||}, \qquad i = 1, 2, \ldots, N-1 \tag{2}$$

where J_0 and J_1 are the coordinates of the torso and the neck joint, respectively [14]. The number of attributes of the feature vector, f, is equal to $2(N-1)$. Considering that in this case $N = 7$, the posture feature vector is composed by 12 attributes, which correspond to the x and y coordinates of the restricted set of joints, excluding the torso which was used as reference.

The signal containing the skeleton features for each frame was segmented by 50 %-overlapping moving windows with a size of 3 seconds as the inertial ones, and for each window the mean x and y joints' coordinates were extracted.

4.3 Features Reduction And Datasets Creation

At the end of the features extraction, a total of 32 features were extracted from inertial sensors (index and finger) and 12 from the skeleton data. The Kruskal Wallis test was applied to obtain the most significant feature vector in distinguishing the group of instances. This test confirmed that the ten gestures, which characterized the activities under investigation, were statistically different for all the above features ($p < 0.05$). Finally, a correlation analysis was performed in order to retain only the significantly uncorrelated features (Correlation Coefficient < 0.85). At the end, the remaining features were combined into different combination of sensors considering also the two cameras' points of view (Tab. 3). By knowing the acquisition frequency of the glove, i.e. 100 Hz, and the one of the cameras, i.e. about 30 frames per second (fps), both incoming data have been synchronised according to the recorded timestamps.

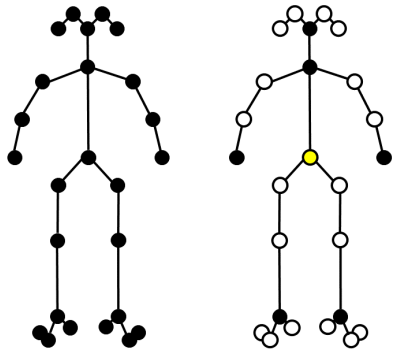

Fig. 4: Representation of the human skeleton: original with 25 joints (left) and subset of selected joints (right) with the torso joint as reference (in yellow).

Table 3: Combination of Sensors.

Acronym	**Combination**
FC	Frontal Camera
LC	Lateral Camera
I	Index finger
IW	Index finger and Wrist
I+FC	Index finger with Frontal Camera
I+LC	Index finger with Lateral Camera
IW+FC	Index finger and Wrist with Frontal Camera
IW+LC	Index finger and Wrist with Lateral Camera

4.4 Classification

The stand-alone systems (i.e. I, I+W, FC, LC) have first been classified to evaluate their performances. Then, all the different combinations, above mentioned, have been classified as detailed in Fig. 5.

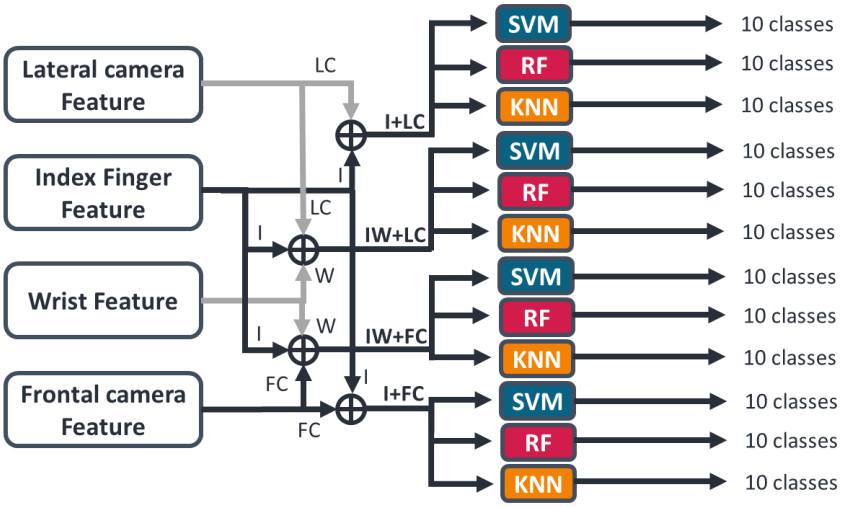

Fig. 5: Feature-level Fusion scheme.

For each classification model, the training set is composed by the 70% of the original dataset, while the test set is composed by the remaining 30%. A 10-fold cross-validation was carried out on the training set. In particular, ten different models were created at the end of the training phase. The 10-fold cross-validation considers the 90% of the initial training set to train the model (train set) and the remaining 10% to evaluate it (validation set). The data in the mentioned sets change at each iteration, to prevent the model from overfitting. The final classification results are based on an average of the performances. Three supervised machine learning algorithms were used in the stand-alone and in the combined classifications, reported in Tab. 3 :

- Multiclass Support Vector Machine (SVM): it exploits the kernel trick to deal with multiclass problems. It maps the input space into a higher dimensional space by using kernels, to make the problem linearly separable, and then it finds the hyperplane that can separate the two classes with the largest margin. In this work, it has been trained by using Sequential Minimal Optimization (SMO).
- Random Forest (RF): it is an ensemble learning method for classification and regression. It operates by constructing a multitude of decision trees at

training time and outputting the class that is the mode of the classes of the individual trees, with the goal of reducing the variance.
– K-Nearest Neighbor (KNN): it is a simple algorithm which stores all available cases and classifies new cases based on a similarity measures, which are distance functions. They are assigned to the most common class among its k nearest neighbors. If k = 1, the object is simply assigned to the class of that single nearest neighbor.

These classifiers were trained to recognize ten classes for each system, which correspond to the ten activities. The classification procedure was implemented and evaluated in MATLAB and the classification performances were evaluated in terms of accuracy, precision, recall and F-measure. In the following, the confusion matrices, corresponding to the configurations with the best accuracy, have been reported to understand the degree of recognition of the different gestures and to evaluate the performances also at gesture level.

5 Results

For the aim of this work, different combinations of sensors were evaluated and three classifiers were applied. The features retained after the feature selection described in Sec. 4.3 are reported in Tab. 4, for each combination. The classification results of the stand-alone system was taken as a gold-standard reference for comparison.

Table 4: Features selected after correlation analysis.

Index+Wrist		**Index**	**Cameras**
Wrist acc. mean	Index acc. mean	Index acc. mean	Head x
Wrist acc. stdev	Index acc. stdev	Index acc. stdev	Head y
Wrist acc. RMS	Index acc. RMS	Index acc. RMS	Neck x
Wrist acc. skewness	Index acc. skewness	Index acc. skewness	Neck y
Wrist acc. kurtosis	Index acc. kurtosis	Index acc. kurtosis	Left hand x
Wrist acc. SMA	Index acc. SMA	Index acc. SMA	Left hand y
Wrist acc. power	Index acc. power	Index acc. power	Right hand x
Wrist ang. vel. mean		Index ang.vel. mean	Right hand y
Wrist ang. vel. stdev		Index ang.vel. stdev	Left foot x
Wrist ang. vel. power		Index ang.vel. power	Left foot y
			Right foot x
			Right foot y

5.1 Stand-Alone Systems

Inertial Sensors The system was evaluated by considering the index sensor (I) and the combination of the index and the wrist sensors (I+W). In the former

case, the activity is described by a total of 10 features (first column of Tab. 4), while in the latter case 17 features are attributed to each activity. The results, shown in Tab. 5, suggest that inertial sensors are quite good in recognizing human gestures, especially when considering I+W combination (up to 86% of accuracy). Indeed, when considering only the inertial sensor on the index, values of 78% are obtained for accuracy, precision, recall and F-measure.

Table 5: Results obtained by inertial sensors

	I+W				**I**			
	Accuracy	Precision	Recall	F-measure	Accuracy	Precision	Recall	F-measure
SVM	0.86	0.86	0.86	0.85	0.78	0.78	0.78	0.78
RF	0.84	0.84	0.84	0.84	0.68	0.70	0.68	0.68
KNN	0.86	0.86	0.86	0.86	0.78	0.76	0.76	0.76

Visual Sensors The feature selection steps returned that all the skeleton features were uncorrelated and relevant for the proposed task. As reported in Tab. 4, a total of 12 features for frontal camera and 12 features for lateral one were used as input to the classifiers. Classification results indicate that the frontal camera is able to recognize the activities with a 95% of accuracy by considering KNN classifier, with respect to the 87% of the lateral camera, as shown in Tab. 6. The results related to the lateral camera are comparable to the ones achieved by inertial sensors alone when considering both wrist and index (86% of accuracy), and higher than the ones obtained by the index sensor alone (78% of accuracy).

Table 6: Results obtained by the cameras

	FC				**LC**			
	Accuracy	Precision	Recall	F-measure	Accuracy	Precision	Recall	F-measure
SVM	0.94	0.93	0.93	0.93	0.84	0.88	0.84	0.85
RF	0.86	0.86	0.86	0.86	0.82	0.82	0.82	0.82
KNN	0.95	0.95	0.95	0.95	0.87	0.92	0.87	0.88

The results reported in Tab. 5 and Tab. 6 indicate that the RF algorithm is the worst among all, being unable to correctly classify the ten activities. On the contrary, SVM and KNN are comparable in the performance when considering inertial sensors alone (86% and 78% of accuracy for I+W and I, respectively). The results obtained by the cameras suggest that KNN classifier achieves the

best performance in classifying skeleton features (95% and 87% of accuracy for frontal and lateral camera, respectively).

5.2 Fusion At Feature-Level

Four combination of sensors were considered. The combination of index-wrist sensor and frontal camera (IW+FC) is characterized by a total of 29 features, while the combination of the index sensor and the frontal camera (I+FC) describes the activity with 22 features (see Tab. 4). The same number of features characterized the combinations made by substituting the frontal camera with the lateral one. The results achieved by the feature-level fusion are shown in Tab. 7 and Tab. 8.

Table 7: Fusion at Feature-level's Results with Frontal Camera

	IW+FC				**I+FC**			
	Accuracy	Precision	Recall	F-measure	Accuracy	Precision	Recall	F-measure
SVM	0.90	0.90	0.90	0.90	0.95	0.95	0.95	0.95
RF	0.94	0.94	0.94	0.94	0.93	0.93	0.93	0.93
KNN	0.94	0.94	0.94	0.94	0.94	0.95	0.95	0.95

Table 8: Fusion at Feature-level's Results with Lateral Camera

	IW+LC				**I+LC**			
	Accuracy	Precision	Recall	F-measure	Accuracy	Precision	Recall	F-measure
SVM	0.84	0.87	0.84	0.84	0.85	0.90	0.85	0.86
RF	0.96	0.97	0.96	0.96	0.94	0.94	0.94	0.93
KNN	0.87	0.91	0.86	0.87	0.85	0.90	0.85	0.86

The classification performances obtained by the frontal camera and the inertial features are comparable to the ones achieved by the frontal camera classifier (accuracy is 95% and 94%, respectively). An improvement in the performances is obtained by combining the lateral camera with the inertial sensor. The lateral camera (LC) alone got an accuracy up to 87% (see Tab. 6. It improves up to the 96% when considering the lateral camera together with the inertial features (IW+LC). In the same way, precision, recall and F-measure improve up to 97%, 96% and 96%, respectively. Looking at the results obtained by using only the index sensor, the results are prominent. In the combination with the frontal camera (I+FC), the system obtains 95% of accuracy, precision, recall

and F-measure, which slightly outperforms the IW+FC configuration (94% of accuracy, precision and recall). When considering the combination of the index sensor with the lateral camera (I+LC), 94% of accuracy is obtained, comparable to the 96% achieved when considering also the wrist sensor (IW+LC).

Comparing the performances of the classifiers on the fusion at feature-level, the results highlight a general good performance. In the IW+LC and I+LC cases, the RF got an higher accuracy (improved by 9%) with respect to the other algorithms. Both RF and KNN obtain the same performances in the IW+FC combination (94% of accuracy, precision, recall and F-measure), while SVM outperforms in the I+FC combination (95% of accuracy, precision, recall and F-measure).

To analyse the performances of the SVM classification at activity level, the normalized confusion matrices of I+FC and I+LC configurations are shown in Fig. 6 and Fig. 7, in which each cell value has been normalized by the number of observations that has the same predicted class. The system composed by index and frontal camera (see Fig. 6) can correctly recognize the activities 'Talk on the phone (TP)' and 'Brush teeth (BT)', while it encounters some difficulties in recognizing the activities 'Relax on the couch (RC)' and 'Sweep with the broom (SB)'. The confusion matrix in Fig. 7 displays the performances of the I+LC configuration. With respect to I+FC, the SVM classifier is not able to correctly recognize the activity 'Brush teeth (BT)' when considering the lateral camera. In this case, the best recognized activity is 'Eat with the fork (EF)', and good results are obtained also for 'Relax on the couch (RC)' and 'Talk on the phone (TP)' activities.

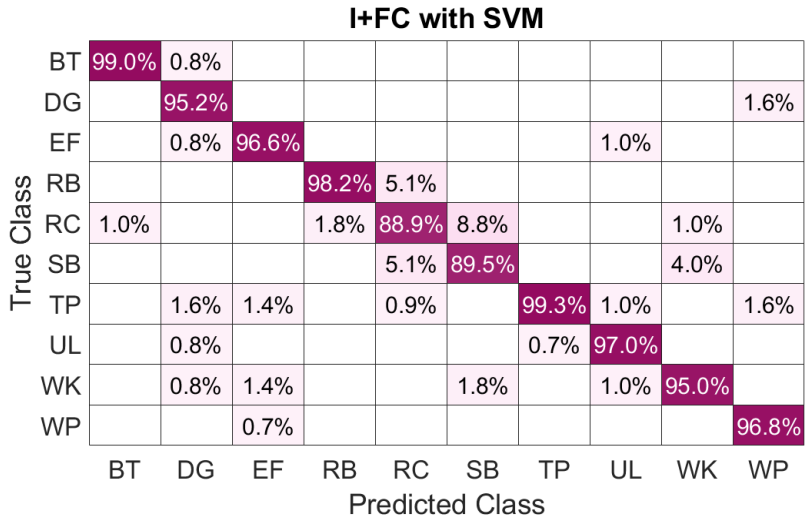

Fig. 6: Normalized confusion matrix obtained by SVM classifier for index and frontal camera

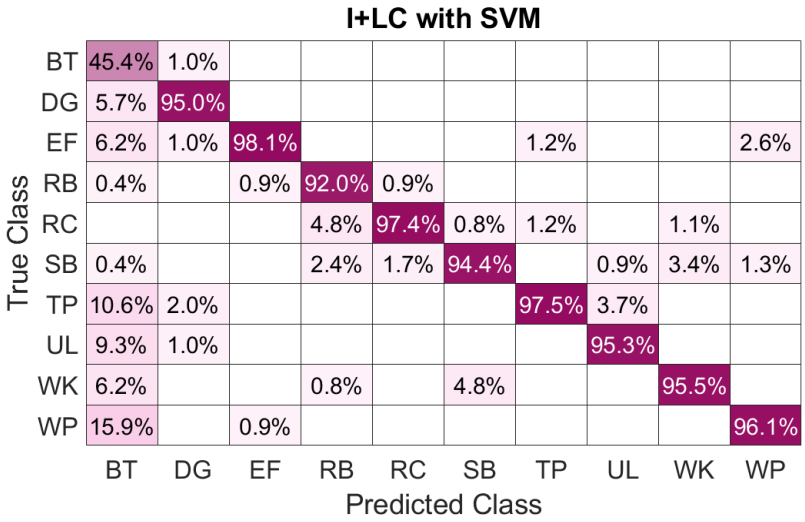

Fig. 7: Normalized confusion matrix obtained by SVM and classifier for index and lateral camera

The data analysis show that the average SUS score was equal to 72.4 (Standard Deviation equal to 14.5), meaning that usability is good (grade B).

6 Discussion And Conclusion

In this work, cameras and wearable inertial sensors have been combined to enhance the capabilities of the robot to recognize human activities. One of the AAL aims is to develop a social robot which is able to recognize human gestures in a non invasive way (i.e. by only exploiting the visual information). The system employed in this work focused on life-like situations, where the users were free to perform the activities. Particularly, in this paper, two different visual perspectives (frontal and lateral) were introduced in the experimental session to explore how the relative position between the robot and the user can affect the recognition task. This work shows that high levels of accuracy are obtained when considering the frontal camera alone. This suggests that the robot can properly recognize the human activities if it is in front of the person. However, this is not a realistic situation. It is quite unlikely that the robot is always perfectly facing the subject. It will more likely be positioned slightly to the side, decreasing its recognition abilities due to occlusion's problem. Moreover, in a life-like situation, the robot and the person could be in relative movement, leading to a decrease in the recognition performances. It is expected that the combination of visual sensors with inertial ones could limit these issues and greatly improve the performances, making the system able to monitor quite well the person in real-time in almost every scenario, avoiding as much as possible delays or mistakes that could affect elderly quality of life.

Four different configurations have been tested with a feature-level fusion approach, i.e. features from frontal camera, wrist and index (IW+FC), from frontal camera and wrist (I+FC), from lateral camera, wrist and index (IW+LC) and from lateral camera and index (I+LC), to understand which is the best combination of sensors.

The selection of gesture appropriately created is made of activities in which the hands are often moved to the head to perform the action, i.e. eating with the fork, drinking from a glass, brush teeth and talk on the phone. These actions are quite similar and difficult to recognize. However, by looking at the confusion matrices, it is possible to appreciate in which extent the system is able to recognize each of them: all the activities are well differentiated and there is no activity which is significantly exchanged for another one. These results suggest that the proposed multi-modal approach could overcome some limitations related to the camera occlusion and similarity between fine gestures.

The results obtained by this fusion of sensors suggested that inertial sensors need to be worn by the user to obtain the best possible gesture recognition. This implies that the person should wear at least two sensors (e.g. smart bracelets and smart ring) during his daily living activities to obtain a good recognition accuracy. There are limitations related to that, because the system could be cumbersome and not easy to wear and to use, especially for elderly people. It is important that they can perform all the movements, which could already be impaired due to their ages, with as little encumbrance as possible. Comparing the performances of the combined classifiers, the results obtained when considering only index and camera are almost the same, and sometimes even better, with respect to the ones achieved by wrist, index and camera combination.

The results achieved in this work outperform the ones obtained by Manzi et al. [14]. In this work, accuracy levels up to 95% are obtained when considering I+FC combination, compared to the 71% achieved by Manzi et al. when considering skeleton data combined to location and wrist features. For this reason, it can be concluded that the use of an inertial ring on the index can be enough to recognise daily living activities and it can also be less bulky. The proposed system could easily become usable in different conditions, since the whole system can adapt to various situations: the robot can be positioned wherever in the room and the wearable sensors can be used everywhere.

In this work, the analysis of the data has been conducted offline. However, this application aims to achieve a system which is able to recognize the gestures in real-time by exploiting the combination of the two sensor modalities.

In conclusion, recognition of daily activities is crucial when monitoring elderly people at home. This is an additional challenge, as the accuracy, precision and usability of obtained data should be high enough to allow the caregivers to remotely monitor the patients, in particular people living with major neurocognitive disorders, allowing them to stay longer at their own place. In the proposed work, the experimentation has been carried out with young healthy people to evaluate whether the system composed by a camera and an inertial ring on the index finger could be used to recognise significant daily gestures, obtaining

positive and promising results. However, it is necessary to test the system also with elderly people to check the performances of the same configuration of sensors. Hence, future experimentation will involve elderly people who could have physical impairment, linked to neurodegenerative diseases like Parkinson and Alzheimer Diseases.

References

1. Measuring usability with the system usability scale. https://measuringu.com/sus/
2. Pepper, soft bank robotics. https://www.softbankrobotics.com/emea/en/pepper
3. Akkaladevi, S.C., Heindl, C.: Action recognition for human robot interaction in industrial applications pp. 94–99 (2015)
4. Bonaccorsi, M., Fiorini, L., Cavallo, F., Esposito, R., Dario, P.: Design of cloud robotic services for senior citizens to improve independent living and personal health management. In: Ambient Assisted Living, pp. 465–475. Springer (2015)
5. Cao, Z., Hidalgo Martinez, G., Simon, T., Wei, S., Sheikh, Y.A.: Openpose: Realtime multi-person 2d pose estimation using part affinity fields. IEEE Transactions on Pattern Analysis and Machine Intelligence (2019)
6. Chen, C., Jafari, R., Kehtarnavaz, N.: A survey of depth and inertial sensor fusion for human action recognition. Multimedia Tools and Applications (12 2015). https://doi.org/10.1007/s11042-015-3177-1
7. Dautenhahn, K., Woods, S., Kaouri, C., Walters, M.L., Koay, K.L., Werry, I.: What is a robot companion-friend, assistant or butler? In: 2005 IEEE/RSJ international conference on intelligent robots and systems. pp. 1192–1197. IEEE (2005)
8. Dawar, N., Chen, C., Jafari, R., Kehtarnavaz, N.: Real-time continuous action detection and recognition using depth images and inertial signals. IEEE International Symposium on Industrial Electronics pp. 1342–1347 (2017). https://doi.org/10.1109/ISIE.2017.8001440
9. D'Onofrio, G., Fiorini, L., Hoshino, H., Matsumori, A., Okabe, Y., Tsukamoto, M., Limosani, R., Vitanza, A., Greco, F., Greco, A., et al.: Assistive robots for socialization in elderly people: results pertaining to the needs of the users. Aging clinical and experimental research **31**(9), 1313–1329 (2019)
10. Fiorini, L., Bonaccorsi, M., Betti, S., Esposito, D., Cavallo, F.: Combining wearable physiological and inertial sensors with indoor user localization network to enhance activity recognition. Journal of Ambient Intelligence and Smart Environments **10**(4), 345–357 (2018). https://doi.org/10.3233/AIS-180493
11. Junker, H., Amft, O., Lukowicz, P., Tröster, G.: Gesture spotting with body-worn inertial sensors to detect user activities. Pattern Recognition **41**(6), 2010 – 2024 (2008). https://doi.org/https://doi.org/10.1016/j.patcog.2007.11.016, http://www.sciencedirect.com/science/article/pii/S0031320307005110
12. Li, H., Shrestha, A., Fioranelli, F., Le Kernec, J., Heidari, H., Pepa, M., Cippitelli, E., Gambi, E., Spinsante, S.: Multisensor data fusion for human activities classification and fall detection. Proceedings of IEEE Sensors **2017-Decem**, 1–3 (2017). https://doi.org/10.1109/ICSENS.2017.8234179
13. Manzi, A., Cavallo, F., Dario, P.: A 3d human posture approach for activity recognition based on depth camera **9914**, 432–447 (11 2016). https://doi.org/10.1007/978-3-319-48881-$3_3$0
14. Manzi, A., Dario, P., Cavallo, F.: A human activity recognition system based on dynamic clustering of skeleton data. Sensors (Switzerland) **17**(5) (2017). https://doi.org/10.3390/s17051100

15. Manzi, A., Moschetti, A., Limosani, R., Fiorini, L., Cavallo, F.: Enhancing Activity Recognition of Self-Localized Robot Through Depth Camera and Wearable Sensors. IEEE Sensors Journal **18**(22), 9324–9331 (2018). https://doi.org/10.1109/JSEN.2018.2869807
16. Matarić, M.J., Scassellati, B.: Socially assistive robotics. In: Springer handbook of robotics, pp. 1973–1994. Springer (2016)
17. Moschetti, A., Fiorini, L., Esposito, D., Dario, P., Cavallo, F.: Recognition of daily gestures with wearable inertial rings and bracelets. Sensors (Switzerland) **16**(8) (2016). https://doi.org/10.3390/s16081341
18. Moschetti, A., Fiorini, L., Esposito, D., Dario, P., Cavallo, F.: Daily activity recognition with inertial ring and bracelet: An unsupervised approach. Proceedings - IEEE International Conference on Robotics and Automation pp. 3250–3255 (2017). https://doi.org/10.1109/ICRA.2017.7989370
19. Reig, S., Forlizzi, J., Steinfeld, A.: Leveraging robot embodiment to facilitate trust and smoothness. In: 2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI). pp. 742–744. IEEE (2019)
20. Rovini, E., Maremmani, C., Moschetti, A., Esposito, D., Cavallo, F.: Comparative Motor Pre-clinical Assessment in Parkinson's Disease Using Supervised Machine Learning Approaches. Annals of Biomedical Engineering **46**(12), 2057–2068 (2018). https://doi.org/10.1007/s10439-018-2104-9
21. Voigt, P., Budde, M., Pescara, E., Fujimoto, M., Yasumoto, K., Beigl, M.: Feasibility of human activity recognition using wearable depth cameras. Proceedings - International Symposium on Wearable Computers, ISWC pp. 92–95 (2018). https://doi.org/10.1145/3267242.3267276
22. Vrigkas, M., Nikou, C., Kakadiaris, I.A.: A review of human activity recognition methods. Frontiers in Robotics and AI **2**, 28 (2015). https://doi.org/10.3389/frobt.2015.00028, https://www.frontiersin.org/article/10.3389/frobt.2015.00028