Enhancing activity recognition of self-localized robot through depth camera and wearable sensors

Alessandro Manzi [†], Alessandra Moschetti [†], Raffaele Limosani, Laura Fiorini, and Filippo Cavallo*, *Member, IEEE*

Abstract— **Robots will become part of our everyday life as helpers and companions, sharing the environment with us. Thus robots should become social and able to naturally interact with the users. Recognizing human activities and behaviors will enhance the capabilities of the robot to plan an appropriate action and tailor the approach according to what the user is doing. Therefore, this paper addresses the problem of providing mobile robots with the ability to recognize common daily activities. The fusion of heterogeneous data gathered by multiple sensing strategies, namely wearable inertial sensors, depth camera, and location features, is proposed to improve the recognition of human activity. In particular, the proposed work aims to recognize ten activities using data from a depth camera mounted on a mobile robot able to self- localize in the environment and from customized sensors worn on the hand. Twenty users were asked to perform the selected activities in two different relative positions between them and the robot, while the robot was moving. The analysis was carried out considering different combinations of sensors to evaluate how the fusion of the different technologies improve the recognition abilities. The results show an improvement of 13% in the F-measure when different sensors are considered with respect to the use of the sensors of the robot. In particular, the system is able to recognize not only the performed activity, but also the relative position, enhancing the robot capabilities to interact with the users.**

Index Terms— **Sensor fusion, wearable sensors, activity recognition, depth camera.**

I. Introduction

In future scenarios, robots will permeate our daily lives, sharing environments such as houses, streets, and offices. Mobile robots can help and support people, but also socially interact with them [1]. In this context, it is useful for a robot to automatically recognize the activities and intentions of humans to cooperate effectively with them. A robot that can understand human activities is capable to plan appropriate reactions according to the situation. These abilities may then lead to more complex robotic services in the fields of security, surveillance, and assistance of elderly people [2]. In addition, enabling the robot to recognize common human activities will enhance the human-robot interaction, helping the robot to understand whether to interact or not with the user and the way to do it. In this way, the interactions would be led by the current activity done by humans. Also understanding the relative position between the robot and the user would influence the human-robot interaction allowing to tailor the approaches based both on the position and on the activities of the person [3].

In the last years, significant research efforts have focused on activity recognition, in particular, using two approaches, namely with external sensors and wearable sensors [4]. The former can include the use of sensors placed in the environment or directly on the objects, whereas the latter concerns devices placed on human body. Smart homes and cameras are example of external sensors. In the first case, the recognition is based on sensors placed on the objects used during the activities, requiring therefore a huge amount of sensors and an update of the system when a new object is added [5]. In the second case, video analysis is used to recognize the activities, in particular, skeleton extraction is used when depth-cameras are adopted [6]. This kind of sensors, however, are linked to issues like privacy, complexity and pervasiveness, which arise from the limitation of the field of view of the camera. The user have to stay in front of the camera to make the system recognize the activity he/ she is performing, limiting the person during his daily life [7].

In order to overcome these problems, according to the state of the art, several strategies were developed and tested. For instance, sensors are used to perceive how the body movements change the electromagnetic noise coming from power lines and electronic devices [8]. In other works, textile capacitive sensors are used to measure the capacitance under the electrode to recognize the activities that are performed [9].

Thanks to the miniaturization and the inclusion of sensors in common wearable objects (i.e. smartwatches, smart wristbands [10]), the use of inertial wearable sensors, for daily activity recognition has been analyzed in several works. 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 [7]. Moreover, wearable sensors are not affected from

* Corresponding author;

[†] A. Manzi and A. Moschetti contributed equally.

A.Manzi, A.Moschetti, L.Limosani, L. Fiorini and F. Cavallo are with the BioRobotics Institute, Scuola Superiore Sant'Anna, Viale Rinaldo Piaggio, 34, 56025, Pontedera (PI), Italy; (Contact Information: +39-0587-672-152; Corresponding author email: filippo.cavallo@santannapisa.it).

TABLE 1
REVIEW OF STUDIES ON ACTIVITY RECOGNITION (DT= DECISION TREE, MLP= MULTI-LAYER PERCEPTRON, RF=RANDOM FOREST, SVM= SUPPORT VECTOR MACHINE, KNN= K-NEAREST NEIGHBORS, HMM= HIDDEN MARKOV MODEL)

Ref.	Sensor Position	Activities	Machine Learning	Results
[14]	Static RGB-Depth camera Accelerometers placed on the wrist and on the waist	13 full body daily activities (database [7])	SVM	Recognition rate: only accelerometers: 56.57% only visual features: 71.52% fusing information: 73.99%
[13]	Static Kinect camera in front of the subject Accelerometers on the wrist and opposite hip	11 actions: general activities that involve both arms or full body activities (Berkeley-MHAD dataset)	SVM, Sparse and Collaborative Representation Classifier, kNN, HMM Feature-level and Decision-level fusion approach	Recognition Rate: only Kinect: > 63.81% both accelerometers: >86% Feature-level fusion approach: 99.13% Decision-level fusion approach: 98.87%
[18]	Static Kinect in front of the user Inertial sensors placed on the wrist or on the thigh according to the action	27 daily activities (UTD-MHAD dataset): 21 actions involving arms with inertial sensor on right wrist and 6 full body activities with inertial sensor on right thigh	Collaborative Representation Classifier Feature-level fusion approach	Accuracy = 79:1% in the fusion approach
[19]	Static Kinect Inertial sensors on the wrists, ankles and back.	11 daily activities: Read, Sleep, Sit idle, Dress, Undress, Brush teeth, Clean a table, Work at the computer, Tidy up the wardrobe, Pick something up, and Sweep the floor	Multiple MLP kNN	F1 score median value = 0:75 (min. median value for sensors fusion approach)
[20]	Static Kinect in front of the user Inertial sensor on the wrist	Ten single hand gestures	multi-HMM Decision-level fusion approach	Recognition rate = 91%
[21]	Static Kinect in front of the user CyberGlove	12 upper body gestures: be confident, have question, object, praise, stop, succeed, shake hand, weakly agree, call, drink, read, and write	Energy-based LMNN classifier	Accuracy = 91% (min. value)
Our work	Depth camera on a moving robot 2 relative positions between the robot and the user IMUs on the wrist and on the index finger	10 common daily activities in 3 different rooms	SVM and RF Decision-level fusion approach	Accuracy = 77% in the best configuration, namely fusion of depth camera, IMUs on wrist and index finger and location (see first row of Table 4)

problems related to illumination variations, background change and body occlusions, but several sensors should be used to recognize whole body movements [11]. Recent studies have demonstrated that the simultaneous use of inertial sensors and depth cameras can improve the recognition of daily activities, merging the advantages of heterogeneous sensors, while compensating for the limitations [12]–[14].

In this context, aim of this work is to combine data from inertial sensors worn by the user, from a depth camera mounted on a mobile robot, and the information about user location in the house given by the robot to recognize ten different activities. The use of this combination of sensors enables, therefore, to increase the pervasiveness and the accuracy of the system to recognize the activities. Furthermore, in addition to getting information about fine movements, wearable sensors allow to perceive the user movements even when the person is not in the field of view of the robot. On the other hand, the use of the depth camera mounted on the robot gives information about the whole body posture, increasing thus the ability to distinguish among different activities. Moreover, information from the depth camera allows to understand the relative position between the user and the robot, enabling it to tailor the approaches according to the relative position. The information obtained by the system (i.e. the activity of the user and the relative position) could be used to adapt the behavior of a mobile robot platform, as expressed and remarked in [3], [15]. The merging of the information enhance the ability of the robot: even when sharing the same environment, the user can be out of the field of view of the robot and other complications may occur (as occlusions of key elements of user gestures), but the sensors worn by the person allow to have information about the activities carried out. The integration of Robotics, Internet of Things and Artificial Intelligence is, therefore, an interesting approach, called “Internet of Robotic Things”, which gives the possibility to design and develop new frontiers in human-robot interaction, collaborative robotics, cognitive robotics, etc. [16].

The rest of the paper is organized as follows. In Section 2 we describe the related works, while the system used and the experimental setup, and the methodology are described respectively in Section 3 and 4. Section 5 provides results and Section 6 a discussion about them. Finally, in Section 6 conclusions are described.

II. RELATED WORKS

As can be seen from literature evidence (see Table 1), different works have analyzed the combination of depth cameras and inertial wearable sensors. Among the works found in literature, Tao et al. [14] presented a comparative study on a database [17] of 13 common home actions acquired in a realistic setting. Chen et al. [13] based their study on the Berkeley-MHAD (Multimodal Human Action Database) dataset, which recorded the actions in a controlled setting without the presence of external noise. In [18], the authors analyzed 27 daily activities (UTD-MHAD dataset), while Delachaux et al. [19] proposed a system composed of a fixed Kinect and five Inertial Measurement Units (IMU) to recognize eleven daily activities. Other works focused on the recognition of hand gestures fusing information from a Kinect camera with an inertial sensor on the wrist [20] or a CyberGlove [21]. In particular, Xiao et al. [21] recognized twelve upper body gestures to improve the natural interaction with robots. Nevertheless, all these works (see Table 1) were conducted using a fixed camera in a controlled environment, limiting the practical use in real cases. Considering the future presence of robots in daily life, it will be possible to exploit mobile cameras placed on it, instead of placing several cameras in the environment. However, the use of robots introduces additional noises related to moving sensors, which can affect the quality of the images, and human-robot relative positions, which can be sub-optimal for activity recognition tasks.

Therefore, the present work aims to go beyond the state of the art by combining data from a depth camera mounted on a mobile platform, able to self-localize in the environment, and from a custom wearable device (SensHand [22]), equipped with inertial sensors. In particular, differently from the aforementioned works, skeleton data extracted from the RGB-D camera were collected while the robot was continuously moving and combined with data from inertial sensors placed on the user's hand. Furthermore, a localization system, using a static map of the environment, was used to gather the location (kitchen, bedroom, living room) where the activity was performed. Merging these information it is possible to recognize actions, which are a combination of hand gestures and full-body postures, overcoming occlusion problems and using only two IMUs. Hence, this work focuses on the investigation of a multi-modal sensor approach to enhance the recognition of common activities using a robot that could monitor people at home and also tailor its action according to the activity the user is performing.

III. Study Design

In this section, we describe the components of the system and the experimental protocol carried out to test the ability to recognize the chosen activities.

A. System Description

The aim of the system is to exploit multiple sensors strategies to enhance the recognition of common daily activities. The architecture depicted in Fig. 1 consists of two devices:

- Mobile Robot able to safely navigate through an indoor environment and featuring an RGB-D camera that provides human skeleton information;
- SensHand, a custom wearable inertial system, able to provide inertial data (accelerations and angular velocities) from multiple sensors.

The components of the system are integrated through a data processing module that handles communication, and synchronization of data, and the recognition algorithms. The modularity of the system allows the evaluation of the classification performance of different configuration of sensing technologies that were adopted during the analysis.

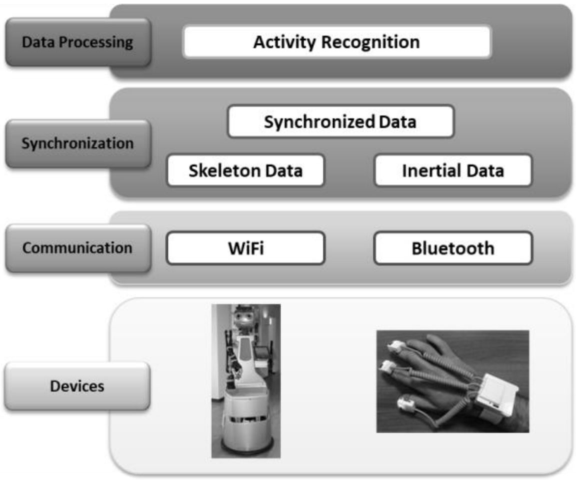


Fig. 1. Overall software architecture of the system.

1) Mobile Robot

The system integrates the DoRo robot ([23]), which is an indoor mobile robot (Fig. 1) , based on the Scitos G5 platform (Metralabs, Germany). The robot module gives both location and skeleton data. In particular, a static map of the home environment was previously built using its laser scanner and a simultaneous localization and mapping (SLAM) based algorithm [24]. The navigation stack, which employs advanced algorithms for obstacle avoidance and path planning included in the commercial Cognidrive software uses the static map and laser scanner data to localize itself within the environment and to avoid obstacles while moving. The navigation system, using the 2-D coordinates of the localization software, infers the area in terms of location (which is limited, in our case, to kitchen, living room, and bedroom). The mobile platform mounts an Asus Xtion depth camera that provides images and depth maps. A software tracker [25] allows the extraction of useful information about the human skeleton. Specifically, it provides a skeleton model of 15 joints, expressed as 3-D Cartesian coordinates, at a sample rate of 10 Hz. Hence, the robot module gives both skeleton and location data.

2) SensHand

The system integrates inertial data acquired with the SensHand (see Fig. 1) , which is a custom device made of four nine-axis inertial sensor units used to perceive the movement of the hand [26]. In particular, the four units are placed on the wrist (similar to current smartwatches) and on three fingers (as rings), typically thumb, and index and middle fingers. The wrist module, which coordinates the other units through the Controller Area Network standard, collects and sends the data to the Control PC via Bluetooth at 50 Hz. On the device, a fourth-order low-pass digital filter with a cutoff frequency of 5 Hz is implemented to remove high-frequency noise. The SensHand has a modular architecture, which allows one or more finger units to be unplugged according to the data required.

TABLE 2
ACTIVITY DESCRIPTION WITHIN LOCATIONS (B: BEDROOM, L: LIVING ROOM, K: KITCHEN).

Activity	Description	Room
CH: Chop	Chop some vegetables continuously	K
DK: Drink with a glass	Take a glass from the table, drink and put it back on the table	B, L, K
EH: Eat with a hand	Take biscuits from a dish and eat them continuously	K
ES: Eat with a spoon	Eat some yogurt with a spoon repeatedly	K
OP: Open a pill container	Open a pill container by unscrewing the cap	B, K
PH: Talk on the phone	Talk on the cellphone	B, L
RD: Read a book	Read a book on the chair (and turn pages)	B, L
RC: Relax on the couch	Sit comfortably on the couch and relax	L
ST: Stir	Stir a liquid in a pot with a wooden spoon continuously-	K
TC: Talk on the couch	Sit on the couch and talk with another person	L

3) *Data Processing Module*

The data processing module is implemented on a PC, which is connected via Bluetooth to the SensHand, and via WiFi to the robot to perform activity classification. In particular, the module retrieves skeleton data from the depth camera mounted on the robot and inertial data from the SensHand. The communication is established using the Robot Operating System (ROS) framework [27]. Since the data are published at different frequencies, an ad-hoc synchronization mechanism was developed. Particularly, for each measurement of the inertial system, poses of skeleton joints were computed as interpolations between the closest data acquired by the depth camera. This procedure considers the relationship between the different coordinate frames and uses the tf library [28] to compute the transformations between frames in time. Regarding the SensHand, the receiving time of data is used as the timestamp of the acquired information; the delay introduced by the Bluetooth communication is thus considered negligible according to this specific application. Moreover, the data processing module runs the activity recognition algorithms used for the classification. Further details on their implementation are given in Section 4.

B. *Experimental Protocol*

Ten activities were chosen from the Cornell Activity Dataset [29], extracting the ones that could be more useful to monitor daily actions considering the correlation between changes in behavior and onset and worsening of cognitive problems in elderly people [30]. Furthermore, two eating activities were added to include also food habits. The Table 2 lists the ten performed activities. Each activity is performed in one or more rooms. Twenty young healthy participants, 8 females and 12 males (19 right-handed and one left-handed), whose ages range from 22 to 37 (29.8 ± 4.2) were involved in the experimental session. The experiments were conducted in the DomoCasa Lab (described in Fig. 2), a fully furnished apartment located in Peccioli, Pisa. This realistic setting was chosen to minimize unnatural movements coming from a laboratory setting. Considering the results obtained in [26], the SensHand was used only with the wrist and index finger sensor units. Users were asked to wear it on the wrist, and on the intermediate phalange of the index finger. Each activity was performed for 1 minute in two different modalities to acquire data from two different points of views. In the first case the action was done in front of the robot, while in the second case, the person performed the activity sideways, thus having the dominant hand on the opposite side with respect to the mobile robot, increasing the complexity for vision-based system due to the occlusion problems. No instructions were given to the users about how to perform the activities, they were, indeed, left free to grab the objects and act in the way they preferred. During the acquisition, each activity was labeled manually by an operator using an ad-hoc web interface.

During the experimentation the robot was continuously moving, hence considering noise linked to the movement of the robot while reaching the person. Therefore, DoRo moved forward and backward continuously between two goal positions using its autonomous navigation system. The two positions were chosen to maintain the user in the field of view of the depth camera, to be able to get the skeleton data during the entire experimental session, as the topic of this work was not related to people detection and tracking.

IV. Activity Recognition

This section describes the activity recognition process, which is the core of the system. It employs the use of supervised machine learning algorithms to perform classification of the activities. The first phase involved the extraction of the features from the sensors, namely the location, skeleton, and inertial data. These features were then used to train two independent classifiers for the skeleton and inertial data. Indeed, a fusion-at-decision-level approach [31] was implemented, and an additional classifier was trained on the outputs of the aforementioned independent classifiers (see Fig. 3). All the processes described in this section were

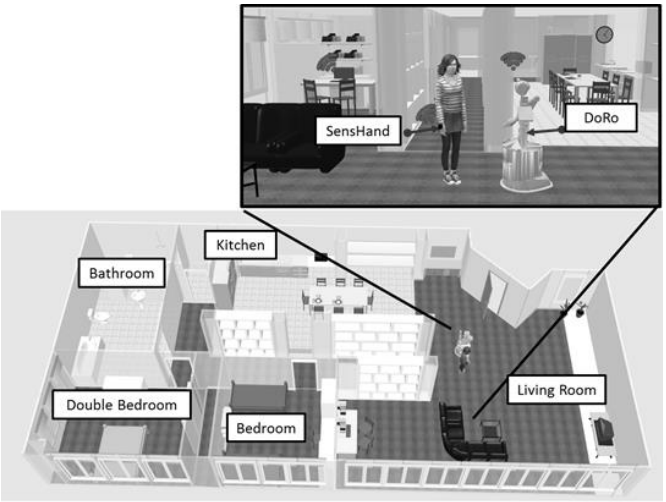


Fig. 2. Representation of the experimental setting. It includes the DoRo robot, and the SensHand. We considered only kitchen, bedroom, and living room.

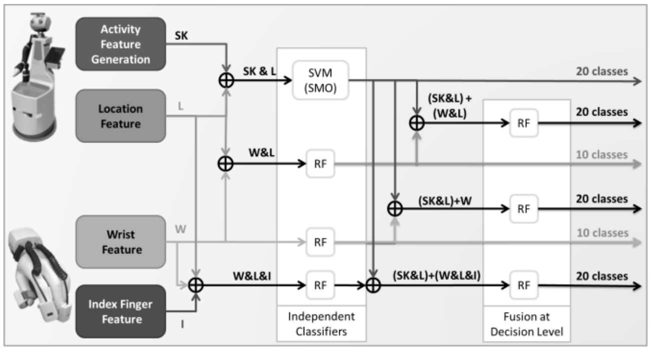


Fig. 3. Software architecture of the activity recognition system. Two independent classifiers are trained on location, skeleton, and inertial data. A decision level fusion approach is used to merge the information.

implemented using Weka Workbench [32].

A. Feature Extraction

Three different types of features are used by the classifiers: location of the user, skeleton activity, and inertial features.

User Location: The location of the user is provided by the navigation module of the mobile platform (see Section 2.1). The localization system of the DoRo robot can infer the current location using the Cartesian coordinates given in the environmental map. In our case, the locations were the kitchen, living room, and bedroom (see Fig. 2).

Skeleton Activity Features: The activity features used by the classifier were extracted from the raw skeleton data after processing steps that involve the use of clustering and supervised machine learning algorithms. The employed method is based on the approach presented in [33]; it describes an activity using several sequences of few basic informative postures. The raw skeleton data describe a human skeleton with 15 joints that are represented as three-dimensional Cartesian coordinates. Only a subset of joints is actually used for the extraction of the features i.e. the head, hands, and feet. These data are normalized on the torso joint frame and scaled with respect to the distance between the neck and the torso. After this normalization step, the X-means clustering algorithm is used to find the centroids of the skeleton, which can be seen as the most informative postures for the input stream. Then, a sequence of posture transition is generated from the original input, considering the obtained clusters. Finally, a sliding window, composed of five elements, is applied to the compressed sequence, generating new instances that represent the activity features for the current input sequence. The choice of these parameters (i.e. the six skeleton joints and the sliding window with a length of five) follows the experiments conducted in [33].

TABLE 3
DESCRIPTION OF THE ADOPTED CONFIGURATIONS.

Configuration	Description
S&L: Skeleton & Location	Classification using only features from the robot, namely skeleton and location.
W: Wrist	Classification using only features from the wrist sensor.
W&L: Wrist & Location	Classification using features from the wrist sensor and location.
(S&L)+W: Skeleton & Location plus Wrist	Decision-level fusion of the classification outcomes obtained independently from the robot and the wrist
(S&L)+(W&L): Skeleton & Location plus Wrist & Location	Decision-level fusion of the classification outcomes obtained independently from the skeleton and the wrist, including location in both cases.
(S\&L)+(I\&W\&L): Skeleton & Location plus Index& Wrist & Location	Decision-level fusion of the classification outcomes obtained independently from the robot, and inertial sensors (wrist and index) including location.

Inertial Features: In this analysis, we considered only the wrist and index finger sensors, according to the results obtained in [26]. Hence, using the acceleration norm, the following features were extracted for both sensors: mean, standard deviation, variance, mean absolute deviation, root mean square, energy, and IAV (integral of the magnitude of the acceleration vector). These features were used to model the whole activity sequence.

B. Classification

To investigate the use of multi-modal sensors in recognizing commons activities, we evaluated our system following six configurations according the considered sensors (see Table 3). In particular, the modularity of the system allowed to test different combination of features starting from data of the independent systems, robot on one side and SensHand on the other one, to the final combination where all the sensors were considered. A leave-one-subject-out cross-validation protocol (LOSO) analysis was carried out, where 19 participants were used for training and the left-out one for testing. The results are, therefore, an average of the performance for each user used as a test set. This type of analysis is useful to assess the recognition ability of the classification system in case of unseen data.

Classification with Skeleton and Localization Data: The location and skeleton activity features extracted as described in Section 3.1.1 and 3.1.2 were used to train a supervised machine learning classifier. Since we were interested in recognizing the relative position of the person, the total number of output classes was 20 (Fig. 3). We adopted a multiclass Support Vector Machine (SVM), trained with Sequential Minimal Optimization (SMO). The multiclass version was implemented by combining several binary SVMs using a one-versus-one strategy.

Activity Classification with Inertial Data: The inertial data were classified by training a random forest; this method provides the best performance with this type of data [34]. Initially, only the features extracted from the wrist sensor were considered (Fig. 3). Next, we included in the classification process the location information as well (Fig. 3). As final setup step, we evaluated the entire system, inputting into the random forest the features extracted from the wrist and index fingers plus the location. In all these cases, the classifier was trained to recognize the classes without the relative position of the user, because the sensor itself does not have the information to discriminate the user position. Hence, the number of activities to be recognized is 10.

Fusion at Decision Level: To combine the outcomes of the two aforementioned independent classifiers, we adopted a fusion-at decision-level scheme [35]. Thus, we trained a random forest that takes as input the outputs of the independent models and classifies the activities in both front and side position (Fig. 3).

TABLE 4
CLASSIFICATION RESULTS ACCORDING TO DIFFERENT FEATURES IN TERMS OF ACCURACY, F-MEASURE, PRECISION AND RECALL WITH LOSO ANALYSIS(S=SKELETON, W=WRIST, L=LOCATION, I=INDEX)

	S&L	W	(S&L)+W	W&L	(S&L)+(W&L)	(S&L)+(I&W&L)
Accuracy	0.64	0.61	0.71	0.72	0.71	0.77
F-measure	0.64	0.54	0.70	0.66	0.69	0.76
Precision	0.64	0.54	0.70	0.66	0.70	0.76
Recall	0.65	0.55	0.70	0.66	0.68	0.75

V. Results

Our results show that the fusion-at-decision-level approach improves the classification accuracy compared to the use of the independent classifier. Table 4 reports the obtained performances in terms of accuracy, F-measure, precision, and recall of all six configurations. Using only the features provided by the robot (i.e. skeleton and location data), the overall accuracy is 0.64 and the F-measure is 0.64. In particular, if we consider both front and side cases, the activities, relax on the couch (RC) and talk on the couch (TC), have good recognition rates (see Fig. 4), while among the lowest are drink (DK), and eat with hand (EH). Also the stirring (ST) activity has low accuracy values because the arm is moving too close to the body, and the classifier has difficulties using only skeleton data.

The classification performance obtained with the wrist features are comparable to the ones of the robot (accuracy equals 0.61 and F-measure equals 0.54). However, these results are not directly comparable since the classifiers trained with inertial data gives as output the activities without considering the user relative position (10 classes instead of 20). At this stage, the location feature is not used, because it is our aim to evaluate the wrist features separately from the robotic system.

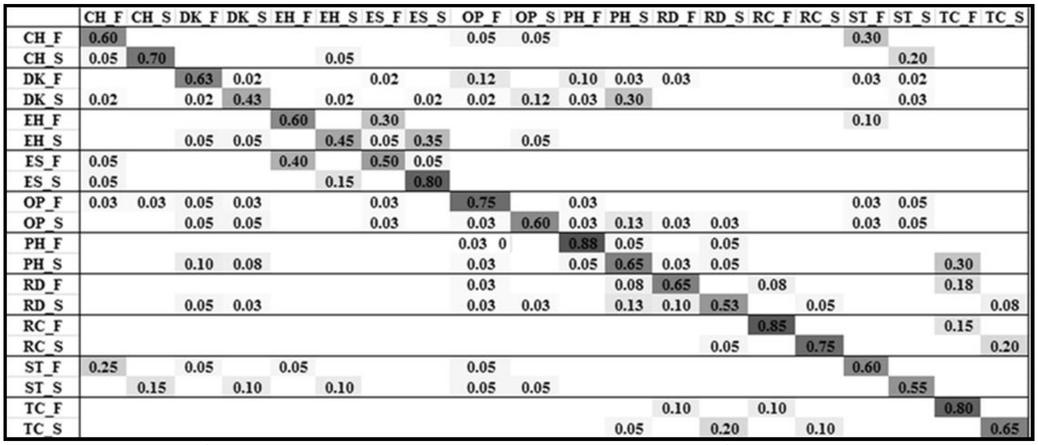
The results obtained combining the above classifiers (i.e. skeleton plus location and wrist) with a decision-level fusion approach lead to an improvement of 7% in accuracy. These results show that the fusion of the two sensors improves the recognition rate of the selected activities. Therefore, on the basis of these results, the location features were added to the wrist ones in the independent classifier. Moreover, in order to consider the robotic system and the SensHand as an unique system, the features were merged at a fusion level (i.e. skeleton plus location and wrist plus location) obtaining accuracy equals to 0.71 and F-measure equals to 0.69.

Finally, our last configuration consists of the introduction of the index finger feature into the system. The final values of accuracy and F-measure are 0.77 and 0.76, respectively. Even if the use of the index sensor implies an increase in the number of sensors to be worn by the user, it improves the F-measure of 13% with respect to the robot feature only and of 6% with respect to the fusion scheme without the index finger, thus justifying the use of an additional sensor.

VI. Discussion

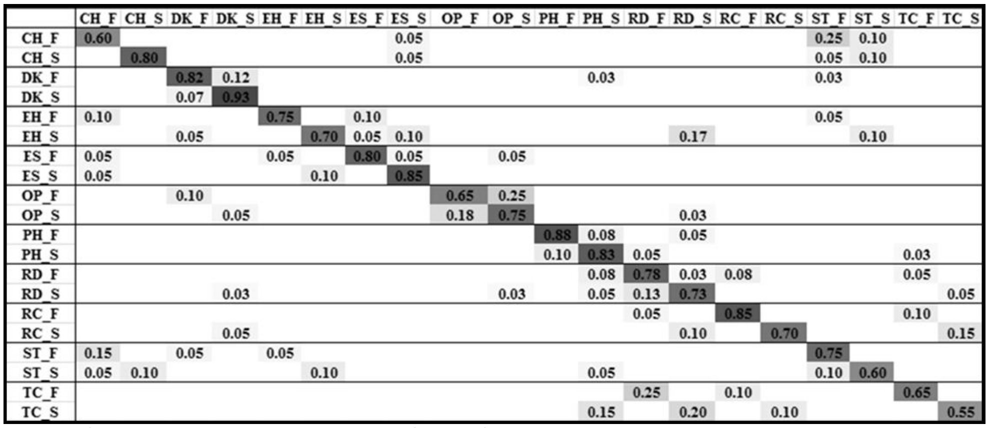
In this work, we proposed a combination of depth camera and wearable sensors to enhance the capability of a robot to recognize some human activities. Six different configurations were tested to recognize ten activities, i.e. features from skeleton and location of the user (robot features), features from the wrist sensor only, fusion at a decision level of features from the robot and from the wrist, features from the wrist and location, fusion at decision level of features from the robot and from the wrist and location, and fusion at decision level of features from the robot and from the wrist and index sensors and location. This last configuration (Fig. 5) increases the recognition rate of drink (DK), talk on the phone (PH), eat with the hand (EH), and stir (ST). Therefore, this final configuration can recognize quite well the proposed activities, even very similar ones such as eat with the hand (EH) and eat with the spoon (ES).

Results show how the proposed system is able to distinguish among daily activities, employing a mobile depth camera and only two inertial sensors, whereas other works are not focused on daily routine activities [13], [18] and use fixed cameras [14] and/or a higher number of inertial sensors [19].



	CH_F	CH_S	DK_F	DK_S	EH_F	EH_S	ES_F	ES_S	OP_F	OP_S	PH_F	PH_S	RD_F	RD_S	RC_F	RC_S	ST_F	ST_S	TC_F	TC_S
CH_F	0.60								0.05	0.05							0.30			
CH_S	0.05	0.70				0.05												0.20		
DK_F			0.63	0.02			0.02		0.12		0.10	0.03	0.03				0.03	0.02		
DK_S	0.02		0.02	0.43		0.02		0.02	0.02	0.12	0.03	0.30						0.03		
EH_F					0.60		0.30										0.10			
EH_S			0.05	0.05		0.45	0.05	0.35		0.05										
ES_F	0.05				0.40		0.50	0.05												
ES_S	0.05					0.15		0.80												
OP_F	0.03	0.03	0.05	0.03			0.03		0.75		0.03						0.03	0.05		
OP_S			0.05	0.05			0.03		0.03	0.60	0.03	0.13	0.03	0.03			0.03	0.05		
PH_F									0.03 0		0.88	0.05		0.05						
PH_S			0.10	0.08					0.03		0.05	0.65	0.03	0.05					0.30	
RD_F									0.03			0.08	0.65		0.08				0.18	
RD_S			0.05	0.03					0.03	0.03		0.13	0.10	0.53		0.05				0.08
RC_F															0.85				0.15	
RC_S														0.05		0.75				0.20
ST_F	0.25		0.05		0.05				0.05								0.60			
ST_S		0.15		0.10		0.10			0.05	0.05								0.55		
TC_F													0.10		0.10				0.80	
TC_S												0.05		0.20		0.10				0.65

Fig. 4. The confusion matrix using only the robot (i.e. skeleton and location features). F= Front, S= Side.



	CH_F	CH_S	DK_F	DK_S	EH_F	EH_S	ES_F	ES_S	OP_F	OP_S	PH_F	PH_S	RD_F	RD_S	RC_F	RC_S	ST_F	ST_S	TC_F	TC_S
CH_F	0.60							0.05									0.25	0.10		
CH_S		0.80						0.05									0.05	0.10		
DK_F			0.82	0.12								0.03					0.03			
DK_S			0.07	0.93																
EH_F	0.10				0.75		0.10										0.05			
EH_S			0.05			0.70	0.05	0.10						0.17				0.10		
ES_F	0.05				0.05		0.80	0.05		0.05										
ES_S	0.05					0.10		0.85												
OP_F			0.10						0.65	0.25										
OP_S				0.05					0.18	0.75				0.03						
PH_F											0.88	0.08		0.05						
PH_S											0.10	0.83	0.05						0.03	
RD_F												0.08	0.78	0.03	0.08				0.05	
RD_S				0.03						0.03		0.05	0.13	0.73						0.05
RC_F													0.05		0.85				0.10	
RC_S				0.05										0.10		0.70				0.15
ST_F	0.15		0.05		0.05												0.75			
ST_S	0.05	0.10				0.10						0.05					0.10	0.60		
TC_F													0.25		0.10				0.65	
TC_S												0.15		0.20		0.10				0.55

Fig. 5. The confusion matrix using the combination of all the features (skeleton, location, wrist, and index). F= Front, S= Side.

Indeed, our dataset was made of activities, which involved the use of the hands, that often were moved to the head to complete the action, e.g. eating or drinking. These actions were quite complicated and difficult for the camera to recognize. Additionally, the activities were recorded both in front of the robot and sideways, occluding the dominant arm; therefore, the problem of occlusion that can affect the depth camera was considered.

In contrast to other works, we mainly focused on real operative conditions, where the users were free to perform the activities in the way he/she preferred.

The system can also distinguish between the different points of view of the robot, and therefore can give information about the position of the user with respect to the robot that could use this detail to approach the user in the best way possible. The robot, moving around the house, could be able to perceive what the user is doing, whether he is eating or drinking or talking to somebody, or simply sitting alone on the sofa becoming bored. In this way, it could approach the user in a proper manner, choosing a tailored way to get closer to the person and to interact with him/her, becoming thus more social. It could also suggest proper things to do according to the activities performed during the day and according to the mood of the persons, trying also to improve it in case the user looks sad or bored.

Thanks to the developed interface, the training dataset was easily created. It was necessary, in fact, for a person to observe the user and simply press a button to label the ongoing activity. Moreover, the proposed system is easily usable in different conditions, since the whole system can adapt to different situation: the robot can simply learn new map and the wearable sensors can be used everywhere.

Even if the use of the SensHand device could be perceived as invasive for daily use, in this experimental work, it was useful to demonstrate the technical feasibility and advantages of integrating hand movement data when the person is not in the field of view of the robot, thus making the system as a whole more ubiquitous. Anyway, due to the limited wearability of SensHand at this stage, during the experimental sessions subjects were always asked if they had some problems with the sensors in terms of obstrusiveness and/or impairments in performing exercises. We are confident that this issue will be soon overcome, thanks to the recent growing interest for smart jewelry [36], where sensors can be part of already used objects, becoming thus more acceptable. Considering this trend, further development of the SensHand is already in progress with a wireless version made of a bracelet and a small ring [37]. In this way, the system can become less invasive and cumbersome and more usable in daily activities. Furthermore, considering the increasing interest in robots as part of daily life, it is reasonable to find a good trade-off between robotic and wearable technologies to exploit the advantages of a heterogeneous system and to improve the abilities of the robot to understand the user activities and adapt its behavior according to the person.

VII. Conclusion

This work demonstrates that the fusion of data coming from a depth camera, placed on a mobile robot, and inertial sensors placed on the users' hands, could improve the recognition accuracy. The ability to distinguish among the activities is improved when the system is considered as a whole, thus complementing the camera with the information about the location of the user and the features of the wrist and index sensor. The movement of the robot and the realistic environment, while adding noise to the acquired data, make the dataset come closer to a real application. Despite the added complexity of considering realistic conditions, these results suggest that multiple sensing strategies can recognize the different activities and the position of the robot with respect to the user, thus enhancing the capabilities of the robot to interact with the user. Future works will explore also other daily activities and more complicated realistic scenarios with the use of more miniaturized ring-shaped wearable sensors.

References

[1] M. Vieira, D. R. Faria, and U. Nunes, "Real-time application for monitoring human daily activity and risk situations in robot-assisted living," in *Robot 2015: Second Iberian Robotics Conference*, 2016, pp. 449–461.

[2] M. Aquilano, F. Cavallo, M. Bonaccorsi, R. Esposito, E. Rovini, M. Filippi, D. Esposito, P. Dario, and M. C. Carrozza ., "Ambient assisted living and ageing: Preliminary results of RITA project," in *Engineering in Medicine and Biology Society (EMBC), 2012 Annual International Conference of the IEEE*, 2012, pp. 5823–5826.

[3] K. L. Koay, D. Syrdal, R. Bormann, J. Saunders, M. L. Walters, and K. Dautenhahn, "Initial Design, Implementation and Technical Evaluation of a Context-aware Proxemics Planner for a Social Robot," in *International Conference on Social Robotics*, 2017, pp. 12–22.

[4] S. C. Mukhopadhyay, “Wearable sensors for human activity monitoring: A review,” *IEEE Sens. J.*, vol. 15, no. 3, pp. 1321–1330, 2015.

[5] A. Wang, G. Chen, J. Yang, S. Zhao, and C.-Y. Chang, “A comparative study on human activity recognition using inertial sensors in a smartphone,” *IEEE Sens. J.*, vol. 16, no. 11, pp. 4566–4578, 2016.

[6] L. Lo Presti and M. La Cascia, “3D skeleton-based human action classification: A survey,” *Pattern Recognit.*, vol. 53, pp. 130–147, 2016.

[7] O. D. Lara and M. A. Labrador, “A survey on human activity recognition using wearable sensors.,” *IEEE Commun. Surv. Tutorials*, vol. 15, no. 3, pp. 1192–1209, 2013.

[8] G. Cohn, D. Morris, S. Patel, and D. Tan, “Humantenna: using the body as an antenna for real-time whole-body interaction,” in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2012, pp. 1901–1910.

[9] J. Cheng, O. Amft, G. Bahle, and P. Lukowicz, “Designing sensitive wearable capacitive sensors for activity recognition,” *IEEE Sens. J.*, vol. 13, no. 10, pp. 3935–3947, 2013.

[10] I. M. Pires, N. M. Garcia, N. Pombo, and F. Flórez-Revuelta, “From data acquisition to data fusion: a comprehensive review and a roadmap for the identification of activities of daily living using mobile devices,” *Sensors*, vol. 16, no. 2, p. 184, 2016.

[11] S. Mitra and T. Acharya, “Gesture recognition: A survey,” *IEEE Trans. Syst. Man, Cybern. Part C (Applications Rev.*, vol. 37, no. 3, pp. 311–324, 2007.

[12] E. Cippitelli, S. Gasparrini, E. Gambi, S. Spinsante, J. Wåhslény, I. Orhany, and T. Lindhy, “Time synchronization and data fusion for RGB-depth cameras and inertial sensors in AAL applications,” in *Communication Workshop (ICCW), 2015 IEEE International Conference on*, 2015, pp. 265–270.

[13] C. Chen, R. Jafari, and N. Kehtarnavaz, “Improving human action recognition using fusion of depth camera and inertial sensors,” *IEEE Trans. Human-Machine Syst.*, vol. 45, no. 1, pp. 51–61, 2015.

[14] L. Tao, T. Burghardt, S. Hannuna, M. Camplani, A. Paiement, D. Damen, M. Mirmehdi, and I. Craddock, “A comparative home activity monitoring study using visual and inertial sensors,” in *E-health Networking, Application & Services (HealthCom), 2015 17th International Conference on*, 2015, pp. 644–647.

[15] K. Charalampous, I. Kostavelis, and A. Gasteratos, “Robot navigation in large-scale social maps: An action recognition approach,” *Expert Syst. Appl.*, vol. 66, pp. 261–273, 2016.

[16] P. P. Ray, “Internet of Robotic Things: Concept, Technologies, and Challenges,” *IEEE Access*, vol. 4, pp. 9489–9500, 2016.

[17] N. Zhu, T. Diethe, M. Camplani, L.Tao, A. Burrows, N. Twomey, D. Kaleshi, M. Mirmehdi, P. Flach, and I. Craddock“Bridging e-health and the internet of things: The sphere project,” *IEEE Intell. Syst.*, vol. 30, no. 4, pp. 39–46, 2015.

[18] C. Chen, R. Jafari, and N. Kehtarnavaz, “Utd-mhad: A multimodal dataset for human action recognition utilizing a depth camera and a wearable inertial sensor,” in *Image Processing (ICIP), 2015 IEEE International Conference on*, 2015, pp. 168–172.

[19] B. Delachaux, J. Rebetez, A. Perez-Uribe, and H. F. S. Mejia, “Indoor activity recognition by combining one-vs.-all neural network classifiers exploiting wearable and depth sensors,” in *International Work-Conference on Artificial Neural Networks*, 2013, pp. 216–223.

[20] K. Liu, C. Chen, R. Jafari, and N. Kehtarnavaz, “Multi-HMM classification for hand gesture recognition using two differing modality sensors,” in *Circuits and Systems Conference (DCAS), 2014 IEEE Dallas*, 2014, pp. 1–4.

[21] Y. Xiao, Z. Zhang, A. Beck, J. Yuan, and D. Thalmann, “Human--robot interaction by understanding upper body gestures,” *Presence teleoperators virtual Environ.*, vol. 23, no. 2, pp. 133–154, 2014.

[22] F. Cavallo, D. Esposito, E. Rovini, M. Aquilano, M. C. Carrozza, P. Dario, C. Maremmani, and P. Bongioanni, “Preliminary evaluation of SensHand V1 in assessing motor skills performance in Parkinson disease,” in *Rehabilitation Robotics (ICORR), 2013 IEEE International Conference on*, 2013, pp. 1–6.

[23] M. Bonaccorsi, L. Fiorini, F. Cavallo, A. Saffiotti, and P. Dario, “A cloud robotics solution to improve social assistive robots for active and healthy aging,” *Int. J. Soc. Robot.*, vol. 8, no. 3, pp. 393–408, 2016.

[24] G. Grisettiyz, C. Stachniss, and W. Burgard, “Improving Grid-based SLAM with Rao-Blackwellized Particle Filters by Adaptive Proposals and Selective Resampling,” in *Proceedings of the 2005 IEEE International Conference on Robotics and Automation*, 2005, pp. 2432–2437.

[25] J. Shotton, T. Sharp, A. Kipman, A. Fitzgibbon, M. Finocchio, A. Blake, M. Cook, and R. Moore, “Real-time human pose recognition in parts from single depth images,” *Commun. ACM*, vol. 56, no. 1, pp. 116–124, 2013.

[26] A. Moschetti, L. Fiorini, D. Esposito, P. Dario, and F. Cavallo, “Recognition of daily gestures with wearable inertial rings and bracelets,” *Sensors*, vol. 16, no. 8, p. 1341, 2016.

[27] M. Quigley, K. Conley, B. Gerkey, J. Faust, T. Foote, J. Leibs, R. Wheeler, A.Y. Ng., “ROS: an open-source Robot Operating System,” in *ICRA workshop on open source software*, 2009, vol. 3, no. 3.2, p. 5.

[28] T. Foote, “tf: The transform library,” in *Technologies for Practical Robot Applications (TePRA), 2013 IEEE International Conference on*, 2013, pp. 1–6.

[29] J. Sung, C. Ponce, B. Selman, and A. Saxena, “Unstructured human activity detection from rgbd images,” in *Robotics and Automation (ICRA), 2012 IEEE International Conference on*, 2012, pp. 842–849.

[30] D. J. Cook and N. C. Krishnan, *Activity learning: discovering, recognizing, and predicting human behavior from sensor data*. John Wiley & Sons, 2015.

[31] C. Chen, R. Jafari, and N. Kehtarnavaz, “A survey of depth and inertial sensor fusion for human action recognition,” *Multimed. Tools Appl.*, vol. 76, no. 3, pp. 4405–4425, 2017.

[32] I. H. Witten, E. Frank, M. A. Hall, and C. J. Pal, *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, 2016.

[33] A. Manzi, P. Dario, and F. Cavallo, “A human activity recognition system based on dynamic clustering of skeleton data,” *Sensors*, vol. 17, no. 5, p. 1100, 2017.

[34] M. Nabian, “A Comparative Study on Machine Learning Classification Models for Activity Recognition,” *J Inf. Tech Softw Eng*, vol. 7, no. 209, p. 2, 2017.

[35] R. Gravina, P. Alinia, H. Ghasemzadeh, and G. Fortino, “Multi-sensor fusion in body sensor networks: State-of-the-art and research challenges,” *Inf. Fusion*, vol. 35, pp. 68–80, 2017.

[36] A. L. Ju and M. Spasojevic, “Smart jewelry: The future of mobile user interfaces,” in *Proceedings of the 2015 Workshop on Future Mobile User Interfaces*, 2015, pp. 13–15.

[37] D. Esposito, F. Cavallo, “Preliminary design issues for inertial rings in Ambient Assisted Living applications.” In Proceedings of the 2015 IEEE International Instrumentation and Measurement Technology Conference (I2MTC), Pisa, Italy, 11–14 May; pp. 250–255, 2015.