Gesture recognition through inertial and visual data

Federica G. Cornacchia Loizzo
The BioRobotics Institute
Scuola Superiore Sant'Anna
Pontedera, Italy
f.cornacchialoizzo@santannapisa.it

Laura Fiorini
The BioRobotics Institute
Scuola Superiore Sant'Anna
Pontedera, Italy
laura.fiorini@santannapisa.it

Alessandra Sorrentino
The BioRobotics Insitute
Scuola Superiore Sant'Anna
Pontedera, Italy
alessandra.sorrentino@santannapisa.it

Erika Rovini
The BioRobotics Insitute
Scuola Superiore Sant'Anna
Pontedera, Italy
erika.rovini@santannapisa.it

Alessandro Di Nuovo
Department of Computing
Sheffield Hallam University
Sheffield, United Kingdom
a.dinuovo@shu.ac.uk

Filippo Cavallo
Department of Industrial Engineering
University of Florence
Florence, Italy
filippo.cavallo@unifi.it

***Abstract*—This paper presents a system composed of Pepper robot that supports an RGB-D camera and an inertial device called SensHand. The fusion of multi-modal data aims to improve the recognition of gestures. Twenty users performed five activities of daily livings (i.e. Having Lunch, Personal Hygiene, Working, House Cleaning, Relax). Each activity was composed of at least two "basic" gestures for a total of 10 gestures. The acquisition of visual data was performed laterally and frontally to mimic real conditions. Acquired data were analysed off-line considering different combinations of sensors to evaluate how the sensor fusion approach improves the recognition abilities. The results show a significant improvement of the accuracy when fusing camera and sensors data, i.e. 0.81 for the whole system configuration when the robot is in a frontal position from the user and 0.75 when the robot is in a lateral position.**

***Index Terms*—Gesture recognition, Human-Robot Interaction, Inertial sensor, Social robot**

I. Introduction

Nowadays social robots permeate our daily life such as workplaces and hospitals. They should recognize what a human being is doing to properly react. For this reason, it is very important that social robots gain the ability to distinguish these simple gestures even when part of daily living 'scenes'. According to literature, the most commonly used sensors in gesture recognition are RGB-D cameras and inertial wearable sensors [1]. The former are widely available and cost effective. They provide rich texture information of the scene and they are easy to operate, but they have some limitations related to background clutter, occlusion, camera position, intra-subject variations and space constraints [1]. On the other hand, inertial sensors enable coping with a much wider field of view as well as changing lighting conditions, but sensor drifts may occur during long operation times; moreover, measurements are sensitive to sensor location on the body. Several works, such as [2] and [3], focused on the use of multimodal sensors to perform activity recognition. In our previous work [4] we combined data from a depth camera mounted on a mobile platform, able to self-localize in the environment, and from the inertial wearable device SensHand. Our actual work aims at developing a multi-modal system in which inertial and visual data are combined together to offer a robust human gesture recognition. It goes beyond the state of the art by reproducing a more realistic scenario in which the activities to be recognized are part of continuous scenes.

II. Material and Method

The presented system is composed by the social humanoid robot Pepper, endowed with a camera mounted over its tablet on its chest, a wearable inertial device namely SensHand, capable of acquiring inertial data from human's hand and fingers, and a data processing module to integrate the components of the system.

Twenty healthy people were involved in the experimentation, 13 males and 7 females, aged from 19 to 44 years old; they were all right-handed. The experimentation was designed to reproduce a real case scenario, in which the participants were asked to wear SensHand and to perform five different daily living scenes, each for one minute, in front of Pepper robot. Every scene was composed from two or three "basic" gestures, as follows. Having Lunch (HL): eat with the fork (EF) and drink from a glass (DG); Personal Hygiene (PH): brush teeth (BT) and DG; House Cleaning (HC): walk (WK) and sweep with the broom (SB); Working (WO): use laptop (UL), write on a paper (WP) and talk on the phone (TP); Relax (RE): TP, relax on the couch (RC) and read a book (RB). Pepper robot gave instructions to the subject about how to perform the scene and it could repeat the assignment if the participant did not understand. The session was recorded by two RGB-D Intel RealSense cameras, one mounted over Pepper's tablet and the other one located on the right side of the participant at the same height from the ground, in order to acquire data from two different points of view.

III. Gesture recognition

As concerns RGB images analysis, 25 keypoints' features were estimated thanks to the Openpose software [5]. However, only the x and y coordinates of the most discriminative set of joints, composed by head, neck, hands, feet and torso, were analysed, following the approach reported in [6]. Then, the signal containing the skeleton features was segmented by 50 %-overlapping moving windows with a size of 3 seconds as the inertial ones. This procedure was performed for the frontal (FC) and the lateral (LC) cameras. For what concerns the wearable SensHand glove, only the data coming from the wrist and index finger sensors were used [7]. They were first filtered with a 4th order digital low-pass Butterworth filter, segmented in 3 seconds' windows and then, from each of them, different features were extracted. The final dataset was composed by 10 features related to acceleration values, i.e. mean, standard deviation, variance, mean absolute deviation (MAD), root mean square (RMS), skewness, kurtosis, signal magnitude area (SMA), normalized jerk and power, and 6 features to angular velocities, i.e. mean value, standard deviation, variance, MAD, RMS and power. The Kruskal Wallis test was then applied: it confirmed that the ten gestures, which characterized the activities under investigation, were statistically different for all the extracted features ($p < 0.05$). Then, a correlation analysis was performed in order to retain only the significantly uncorrelated features (Correlation Coefficient < 0.85). The system was first evaluated by considering both inertial and visual sensors as stand-alone systems and then by fusing the two sensor modalities at feature-level for a total of eleven datasets: frontal camera (FC), lateral camera (LC), index finger (I), wrist (W), IW, I+FC, I+LC, W+FC, W+LC, IW+FC and IW+LC. These datasets were classified using a a 10-fold cross-validation technique. The final classification results are obtained as average of the performances of the ten created models. Three supervised machine learning algorithms were employed: Multiclass Support Vector Machine (SVM), considering a third order polynomial kernel function, Random Forest (RF) and K-Nearest Neighbor (KNN), setting the k nearest neighbors = 1.The classification performances were evaluated in terms of accuracy, precision, recall and F-measure.

IV. Results and Discussions

After the feature selection process, 19 inertial features were retained for IW configuration and only 10 for I and W ones. All skeleton features were included in the analysis. 3361 instances were analysed for each combination of data including the FC, while 3213 for the ones including the LC. Whereas, 3213 instances were considered for I and W configuration. The results obtained under the multi-modal datasets are better than those achieved when considering stand-alone configurations and they are comparable with the related works. As concern the stand-alone configurations, the results show that I and W configurations obtains accuracy levels up to 0.55 and 0.57, respectively, with Random Forest, while 0.65 of accuracy is obtained when considering the IW combination with SVM. The frontal camera (FC), which has a good view of the user performing the activity, is able to recognize the gestures with 0.77 of accuracy, recall, F-measure and precision, with the SVM classifier. These values decrease when considering the camera positioned on the side (LC). In this case, the accuracy, recall, F-measure and precision are 0.69, 0.70, 0.71 and 0.73, respectively, with KNN classifier. As concern the multi-modal datasets, the fusion-at-feature-level approach improves the classification accuracy compared to the use of the independent classifiers: the system is able to recognize the ten activities with 0.81 (IW+FC) and 0.75 (W+LC) of accuracy as best configurations (obtained with SVM and RF respectively). In a life-like situation, the robot will never be positioned exactly in front of the person performing the activity, but it will be more likely in a non optimal position, e.g. on the side. In this study, we found out that the system decreases its capabilities by only moving 90 degrees the position of the camera, losing 8% of accuracy when its view is lateral. This is due to occlusion problems that lead to greater difficulty in recognizing gestures. This issue can be overcome by fusing the information that is acquired from the cameras, able to capture the gross motor actions of the body, with the inertial wearable device, able to capture the fine movements of the hand. Furthermore, it is important to remark that the experimental session was conducted by reproducing real operative conditions as far as possible: in each scene, participants could switch freely from one activity to another in one minute. In this amount of time, the cameras and the SensHand recorded visual and inertial data continuously, therefore the acquisition stream was unique for each scene. For this reason, the novelty with respect to Manzi et al. [6], whose system achieved 0.77 of accuracy, is that in our work the data corresponding to the transitions from one activity to the other were present in the acquisition signal, and the system revealed to be good enough in classifying them without losing much in terms of performances (0.81 of accuracy). 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.

References

[1] C. Chen, R. Jafari, and N. Kehtarnavaz, "A survey of depth and inertial sensor fusion for human action recognition," *Multimedia Tools and Applications*, 12 2015.

[2] M. Ehatisham-Ul-Haq, A. Javed, M. A. Azam, H. M. Malik, A. Irtaza, I. H. Lee, and M. T. Mahmood, "Robust Human Activity Recognition Using Multimodal Feature-Level Fusion," *IEEE Access*, vol. 7, pp. 60 736–60 751, 2019.

[3] C. Chen, R. Jafari, and N. Kehtarnavaz, "A real-time human action recognition system using depth and inertial sensor fusion," *IEEE Sensors Journal*, vol. 16, no. 3, pp. 773–781, 2016.

[4] A. Manzi, A. Moschetti, R. Limosani, L. Fiorini, and F. Cavallo, "Enhancing Activity Recognition of Self-Localized Robot Through Depth Camera and Wearable Sensors," *IEEE Sensors Journal*, vol. 18, no. 22, pp. 9324–9331, 2018.

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

[6] A. Manzi, P. Dario, and F. Cavallo, "A human activity recognition system based on dynamic clustering of skeleton data," *Sensors (Switzerland)*, vol. 17, no. 5, 2017.

[7] A. Moschetti, L. Fiorini, D. Esposito, P. Dario, and F. Cavallo, "Recognition of daily gestures with wearable inertial rings and bracelets," *Sensors (Switzerland)*, vol. 16, no. 8, 2016.