2021 I-RIM Conference
October 8-10, Rome, Italy
ISBN: 9788894580525
DOI:10.5281/zenodo.5900573

Recognizing Italian Gestures with Wearable Sensors

Matteo Dicenzi[*1, 3], Linda Lastrico[1, 2], Alessandro Carfí[1],
Alessandra Sciutti[3], Fulvio Mastrogiovanni[1], Francesco Rea[2]

***Abstract*—For humans, gestures are a means of communication. In order to create a more spontaneous interaction between humans and robots, social robots should be able to understand the information we convey with gestures. To this aim, (i) we collected a dataset with 1469 examples of twelve common Italian hand-gestures using a custom-made inertial glove, via experiments organized as human-robot interactions, and (ii) we propose an offline gesture recognition model based on a Long-Short Term Memory (LSTM) Recurrent Neural Network (RNN), which achieved an overall accuracy equal to $87.0 \pm 3.7\%$.**

***Index Terms*—Non-Verbal Communication, Gesture Recognition, Recurrent Neural Network, Inertial Measurement Unit**

I. Introduction

In the past few years, research has shown that social robotics may bring major benefits to the lives of people. Robots could be used in public spaces, education and personal care [1]. In order to be truly effective in the interaction, robots should *understand* human communication signals. In a general sense, human communication exploits different modalities. It can be explicit, if two or more people intend to exchange information [2], or implicit, if a person communicates to others unintentionally, e.g., through eye gaze or body posture [3]. Gestural communication is a common way of expressing ourselves with our body, with the hands playing a preferential role. It is an explicit, culture-oriented form of communication, and Italian hand gestures are a well-known, idiosyncratic, example [4]–[6]. These gestures are characterized by a gesture-meaning pair, which is well defined and understood. However, their specificity notwithstanding, they are a natural and spontaneous form of communication.

In a broad sense, gesture recognition is a process involving two main sub-processes, i.e., the definition of a gesture model and its online implementation. To start with, it is often required to collect human data through sensors. These can be image-based, e.g., depth or stereo cameras, or non-image-based, such as accelerometers or gyroscopes, as in this study. The latter, compared to image-based approaches, are a good choice in terms of portability and size of the data they produce. Moreover, it is important to extract only relevant information, by looking at the portions of data referring to the gestures, i.e., its *detection*. This is usually achieved through automatic gesture segmentation techniques which, for example, apply a threshold on the variations between two consecutive velocities [7]. Building a model usually requires solving a multi-classification problem, often addressed through data-driven approaches. A probabilistic classifier can be defined and trained to classify the detected gestures under a *closed world* assumption: each input gesture is necessarily one of the gestures in the dictionary.

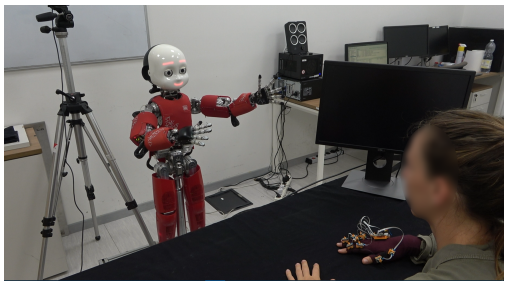

Fig. 1. Data collection setup. iCub asks the participant, wearing the glove on their right hand, to perform the Italian hand gesture that will be shown on the monitor.

In this paper, we address the problem of classifying some of the most typical Italian hand gestures. To this extent, we collected a novel dataset, and developed a recognition architecture capable of classify them *offline*.

II. Methods

In this study, we consider twelve, typically Italian gestures [8]. The complete gesture selection can be seen in the accompanying video[1]. Given the social nature of the study, emphasis must be placed on interactions with participants, an essential step towards replicating gestures. Since the robot can provide the same Human-Robot Interaction (HRI) for everyone, we designed an HRI experiment to collect a gesture dataset, as shown in Figure 1. During the experiment, each volunteer sits in front of a desk and initially holds their hands on the table. For each gesture class, iCub asks the participant to watch a video depicting an example of the gesture. At the end of an emitted sound, the participant can perform the gesture and then return to the initial position. Data collection was carried out using a custom-made data glove, worn by volunteers on their right hand. The glove has two Inertial Measurement Units (IMUs) for each finger. In the thumb, they are close to the metacarpal and intermediate phalanges. In all other fingers, IMUs are always located on the proximal and intermediate phalanges. Moreover, an additional IMU is placed on the hand back, i.e., on the metacarpal bones. Each IMU allows to detect ten features with a 28 Hz frequency: the triaxial

[*] matteo.dicenzi@iit.it
[1] Department of Informatics, Bioengineering, Robotics, and Systems Engineering, University of Genoa, Italy
[2] RBCS, Istituto Italiano di Tecnologia, Genoa, Italy
[3] CONTACT, Istituto Italiano di Tecnologia, Genoa, Italy

[1] Web: https://youtu.be/PFiZEmKKo-Y

linear accelerations, the triaxial angular velocities, and the four orientations expressed as quaternions. The collection process involved 31 Italian volunteers (19 males, 12 females, age: 29 ± 5 years). Each participant experienced the same human-robot interaction, and carried out every gesture four times. Gestures can be represented as a time-series of 110 different features, i.e., 10 features for each of the 11 IMUs. Since each volunteer performed four repetitions of each gesture type, the dataset contains 124 examples for each class and 1.488 examples in total. However, we removed 19 examples (1.27%), performed outside the predetermined time interval (signaled by the described sound). Hence, we considered 1469 examples.

Since raw data contain more than just gesture-related information, we developed an automatic segmentation algorithm to extract (offline) the relevant portions of data from the input sequences. Given a gesture, the algorithm computes, for each IMU, the norm of the acceleration components and identifies the start and the end points of a gesture by applying a threshold estimated when the hand was held steady in the initial rest position. This assumes, obviously enough, that the relevant motion in each trial corresponds to the gesture execution.

The network model we selected is a Long-Short Term Memory (LSTM) Recurrent Neural Network (RNN) for its capability to learn time-dependent information. More specifically, the model is characterized by: an initial masking layer; an LSTM layer, with 200 neurons, "tanh" activation function, "sigmoid" recurrent activation function and $L2$ kernel regularizer; a dropout layer; a dense layer with 200 neurons and a "relu" activation function; a final dense layer with 12 neurons and a "softmax" activation function. The classification model is implemented using Keras and TensorFlow. It is trained with batches of 128 gestures, with (a randomly selected) 70% of the dataset (*training set*) and evaluated, considering k-fold cross-validation approaches, with the rest of the dataset (*test set*). In addition, 20% of the training set (*validation set*) is evaluated for early stopping. Note that the LSTM-model is trained and evaluated offline, providing as inputs complete gestures, i.e. time-series of 160 *timestamps* and 110 *features*. An important consideration regarding the dataset is that each gesture may have a different length. This occurs because each participant carries out the gestures in their own way, taking a slightly different amount of time. For this reason, we pad the data to obtain sequences of 160 samples and use an initial masking layer, which allows us to exclude, from the following computations, the values added during the padding.

III. Results and Conclusions

Figure 2 shows one of the confusion matrices computed when carrying out the k-fold cross validation (kFCV). From the matrix diagonal, we can observe how almost all classes are correctly classified, with a test set accuracy, precision and recall equal to 93%. The overall accuracy, computed as mean of all the kFCV accuracies, is equal to $87.0 \pm 3.7\%$. This level of generalization is a satisfactory result considering two reasons. On the one hand, the number of classes to be recognized is larger if compared to other similar studies [9].

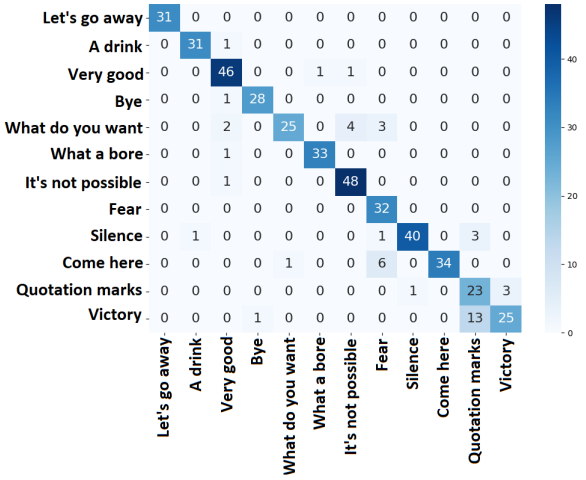

Fig. 2. Confusion matrix summing up the model performance. Gesture labels are on the left/bottom sides. The full label descriptions can be observed in the video (link in note 1).

On the other hand, the variability among gestures performed by people with different experiences does not compromise the generalization capabilities of the model. The main drawback of the current stage of this work is related to the closed world assumption made while training the model: if the input gesture is unknown to the model, it will be anyhow classified as one of the classes in the dictionary. Future developments of this work will be aimed at overcoming this issue. We will consider, for example, an indirect detection module after the classification module [10]. In addition, we plan to release an online version of the architecture, to be used in actual human-robot interaction scenarios.

References

[1] A. Bonarini, "Communication in human-robot interaction," *Current Robotics Reports*, pp. 1–7, 2020.

[2] K. S. Lohan, H. Lehmann, C. Dondrup, F. Broz, and H. Kose, "Enriching the human-robot interaction loop with natural, semantic and symbolic gestures," *Humanoid Robotics: A Reference*, pp. 2199–2219, 2019.

[3] G. Sandini, A. Sciutti, F. Rea, A. Goswami, and P. Vadakkepat, "Movement-based communication for humanoid-human interaction," *Humanoid Robotics: A Reference*, pp. 2169–2197, 2019.

[4] I. Poggi, "Symbolic gestures: The case of the italian gestionary," *Gesture*, vol. 2, no. 1, pp. 71–98, 2002.

[5] J. M. Iverson, O. Capirci, V. Volterra, and S. Goldin-Meadow, "Learning to talk in a gesture-rich world: Early communication in italian vs. american children," *First Language*, vol. 28, no. 2, pp. 164–181, 2008.

[6] A. Kendon, "Gestures as illocutionary and discourse structure markers in southern italian conversation," *Journal of pragmatics*, vol. 23, no. 3, pp. 247–279, 1995.

[7] R. Xie and J. Cao, "Accelerometer-based hand gesture recognition by neural network and similarity matching," *IEEE Sensors Journal*, vol. 16, no. 11, pp. 4537–4545, 2016.

[8] B. Munari, *Supplemento al dizionario italiano*. Muggiani, 1963.

[9] P. Zhu, H. Zhou, S. Cao, P. Yang, and S. Xue, "Control with gestures: A hand gesture recognition system using off-the-shelf smartwatch," in *2018 4th International Conference on Big Data Computing and Communications (BIGCOM)*, 2018, pp. 72–77.

[10] A. Carfi, C. Motolese, B. Bruno, and F. Mastrogiovanni, "Online human gesture recognition using recurrent neural networks and wearable sensors," in *27th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, 2018, pp. 188–195.