The Thirty-Second AAAI Conference
on Artificial Intelligence (AAAI-18)

Learning to Detect Pointing Gestures from Wearable IMUs*

Denis Broggini, Boris Gromov, Alessandro Giusti, Luca Maria Gambardella
Dalle Molle Institute for Artificial Intelligence (IDSIA), USI-SUPSI, Lugano, Switzerland

Abstract

We propose a learning-based system for detecting when a user performs a pointing gesture, using data acquired from IMU sensors, by means of a 1D convolutional neural network. We quantitatively evaluate the resulting detection accuracy, and discuss an application to a human-robot interaction task where pointing gestures are used to guide a quadrotor landing.

Introduction

Hand gestures are an important device that humans use all the time to communicate with each other, and represent an appealing human-robot interaction modality (B. Gromov, L.M. Gambardella, G.A. Di Caro 2016). In this work we want to trigger specific behaviors of the robot(s) with which the user is interacting detecting the moment in which the onset of a pointing gesture occurs, with as little delay as possible. We use as input a stream of acceleration vectors and orientation quaternions values measured by two IMUs, located on the forearm and on the upper arm and transmitting instantaneous values of their accelerometers and gyroscopes at a rate of 50 Hz.

Instead of using manual rules and thresholds to detect this event, we adopt a machine-learning based approach based on a 1D Convolutional Neural Network (Chollet and others 2017) that operates on sequences of samples with a length of 2.5 seconds; a sequence of n consecutive samples covers $n/50$ seconds and is denoted in the following as a window. A window is the input to our classifier, that learns whether it contains the onset of a gesture. The window is classified as positive (1) in case it contains the onset of a pointing gesture, negative (0) otherwise.

With multiple users, we gather a set training and testing datasets using an ad-hoc approach that periodically prompts the users to perform the pointing gesture, and can therefore automatically assign ground truth to every window.

*This project has been partially supported by the Swiss National Center of Competence Research (NCCR) Robotics through the Swiss National Science Foundation

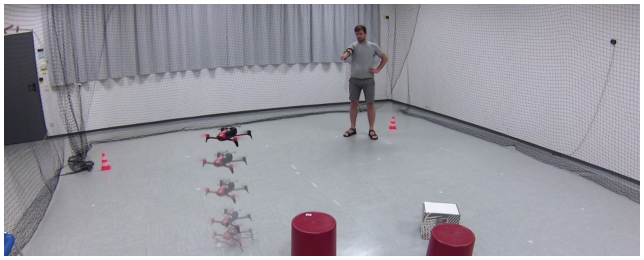

Figure 1: Application to a human-drone interaction task, in which the user can take control of a drone by pointing at it

Dataset Acquisition

Our dataset is a collection of tuples (window, label). In order to acquire a dataset, we recorded several sessions which take place as follows: a user wearing the IMUs performs different movements and actions (such as moving objects, walking around, climbing a ladder). When the system emits a sound (at time t_i^{start}), the actor immediately performs the pointing gesture to an arbitrary spot. A second sound t_i^{stop} indicates the user to stop pointing. The user then continues its previous task.

A recording session contains two time series: the stream of data from the sensors and the set of signals (at known times) that indicate the start and end of each pointing gesture. From a recorded session we can extract many windows: the nth window contains samples from n to $n + 124$.

Now we need to match the collected windows with their label. We consider a reaction time Δr needed by the actor to hear the sound and start performing the pointing, and a time Δx that is the estimated time needed by the actor to completely straighten his arm.

We can use this knowledge to associate each window to the proper label and add the resulting tuple (window, label) into the dataset.

A window that starts before $t_i^{\text{start}} + \Delta r$, ends before $t_i^{\text{stop}} + \Delta r$ and ends after $t_i^{\text{start}} \Delta x$ for *at least* one of the t_i sound pairs is considered a positive (1) window.

On the other hand, a window that starts after $t_i^{\text{stop}} - \Delta r$ or ends before $t_i^{\text{start}} + \Delta r$ for all t_i sound pairs is considered a negative (0) window.

The windows that do not match either of these sets of cri-

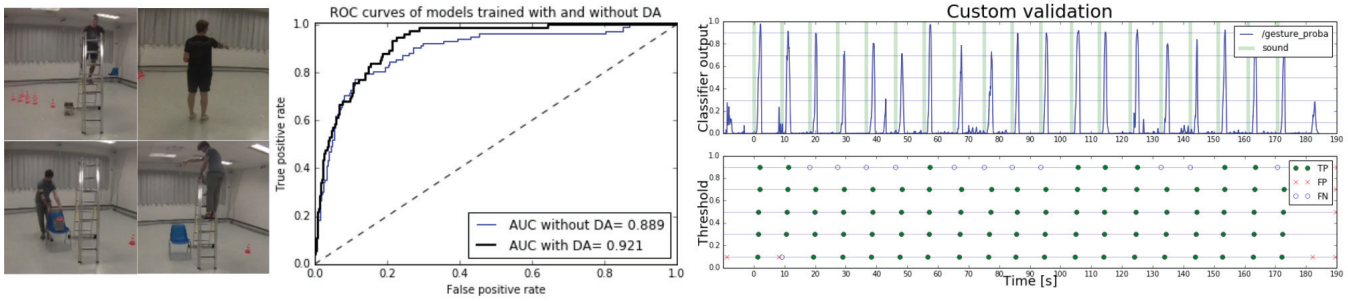

Figure 2: Left: Some recording session scenario. Center: Comparison of two classifiers. Right: Results of the custom validation

teria are considered as ambiguous, as they may only contain a portion of the onset of the pointing gesture, and won't be included in our training dataset.

Data augmentation

The amount of data gathered during our training sessions has been further augmented using data augmentation techniques. When a pointing gesture is performed toward a given direction, the same gesture performed toward a different direction (i.e. a rotation around the world vertical axis) also counts as a pointing gesture. Hence, from a single gesture one can derive any number of other equivalent gestures, by rotating orientation values returned by the IMU by a random amount around the world vertical axis.

Experimental Validation

Dataset

We recorded multiple sessions with a variable duration, recording with different actors. The final dataset contains:

- 12 sessions with a variable duration from 3 to 10 minutes, recorded with 4 different actors, for a total of 85 minutes
- Over 460 different pointing gestures performed
- Approximately 200'000 instances, with over 180'000 instances with label '0' and over 13'000 instances with label '1'

The dataset was split into two chunks: a training set (7 sessions) and the test set (5 sessions). Both the training and the testing set contain gestures performed by the four actors. In this way we are able to evaluate the quality of the model

Performance metrics

We first evaluated the CNN using two different metrics: ROC (Receiver Operating Characteristic), and AUC (Area Under the Curve). Then, we tested the detection system as a whole; ideally, this should trigger an event each time a pointing gesture is performed, and no event otherwise.

We test the detection system as follows. First, we run the detection system on a recorded session that was not included in the training set. Then, we check the outputs of the system according to the validation session: the events triggered by the system between 1 and 3 seconds after a sound are considered *True Positives*, and those triggered outside the small ranges above mentioned are considered *False Positives*. If a sound occurs but no event is triggered then we have a *False Negative*. Finally we score to the system, based on TPR (True Positive Rate) and False Positives per minute. These numbers are a function of the detection threshold.

Results

The trained classifiers yield ROC curves with AUC values over 0.9: we also measured the positive impact of the Data Augmentation strategy we adopted (Figure 2 center).

Then, we have tested our detection system as a whole experimenting with different threshold values. The obtained results are illustrated in the 'Custom Validation' plot, in Figure 2. In the upper plot we show the following data:

- `/gesture_proba`: the outputs of the classifier for a window ending at time t;
- `sound`: the green vertical lines represent the sounds played during the session, where we know that the actor performed the gesture;
- thresholds: the blue horizontal lines represent the different detection thresholds; whenever `gesture_proba` overtakes the threshold, the system detects a gesture and publishes a specific message (detection event).

The bottom plot classifies each event generated by the system: a green circle if the event is correct (TP, True Positive), a red cross if the event is wrong (FP, False Positive) and a blue empty circle if there is a gesture without any event generated (FN, False Negative).

Conclusions

We presented a detection system for pointing gestures based on a 1D CNN operating on IMU data; the approach is effective and has been used for an Human-Drone Interaction task.

References

B. Gromov, L.M. Gambardella, G.A. Di Caro. 2016. wearable multi-modal interface for human multi-robot interaction, in safety, security, and rescue robotics (ssrr). *IEEE International Symposium*.

Chollet, F., et al. 2017. Keras. https://github.com/fchollet/keras.