An Attention System for Social Robot

Abolfazl Zaraki, Daniele Mazzei, Nicole Lazzeri, Danilo De Rossi
Interdepartmental Research Center ‘E. Piaggio’, Faculty of Engineering, University of Pisa, Italy

Why Social Robots Must be Able to Perceive and Analyse the Surrounding Environment?
Thanks to the last progress in biomimetic materials and advances in control techniques social robots are nowadays becoming a reality. This new generation of humanoid robots is growing up and, unlike conventional robots, they are required to have social interaction capabilities to engage with human empathic and emotional relationships [1]. Social robots will be able to act as instructor, assistant, therapist and domestic staff. In order to serve in such positions, they must be able to perceive and analyse the surrounding environment in the way that human being does.

A System that Manipulate Robot ’s Attention Based on the Perceived Stimuli:
As humans do, social robots have to be able to classify objects and people according to different taxonomies and analyse them in detail extracting various properties and distinctive features. Humans must be described using a taxonomy in which various parameters such as ID, position, recognized gesture, facial expression, mouth and eyes state (open/closed), speaking probability, skeleton joints coordinates are included. A system is required to precisely analyse and extract social features from what can be seen by sensors and drive the attention of social robots in human robot interaction.

Material and Method
The general structure of the social attention system is shown in Fig.1. It consists of three layers: acquisition, analysis and fusion which continuously analyse the dynamic 3D environment.

A. Acquisition layer
This layer is devoted to acquire raw data from various sensors. In this preliminary implementation Microsoft Kinect™ sensor has been used as a marker less motion tracking system extracting RGB and depth scene images.

B. Analysis Layer
This layer is the meta-scene creation core. Using Microsoft Kinect Software Development Kit (SDK) together with ad-hoc developed algorithms to transfer data. Collected raw data are analysed extracting the following scene related information: subjects’ position, distance and angle all with respect to the sensor and arrival time. The system extracts the full skeleton information for the two subjects in front of the robot.

In order to analyse the facial features of subjects, SHORE™ (Sophisticated High-speed Object Recognition Engine) is used [2]. SHORE is a facial analysis library able to extract a range of facial features from 2D images: face, eyes, nose and mouth positions, gender classification and facial expressions (“Happy”, “Surprised”, “Angry” and “Sad” as percentages).

C. Fusion Layer
The fusion layer has two main functions: pre-processed data analysis and calculation of parameters that can’t be directly inferred from single sensors raw data. In this layer the core of the attention systems is also implemented: the score calculation algorithm. The main task of this algorithm is to select a target among subjects on which FACE has to focus its attention. In order to select a subject, the attention score is calculated using Eq. (1). Inspired by human being attention system, each stimulus in the scene drives the attention of the system on the base of its social relevance. As an example someone who talks to the robot or shows an angry face will obtain a higher score.

References:

In this preliminary integration test \( F_i \) functions were set as linear and the \( K_i \) as reported:

\[
\begin{align*}
K_1 & = [K_{11}, K_{12}, K_{13}, K_{14}] = [1, 4, 3, 2, 1] \\
K_2 & = 3 \\
K_3 & = 2 \\
K_4 & = 1
\end{align*}
\]  (2)

The \( K_i \) were selected in order to mimic an attention system with the following stimuli priority: arrival time, speaking probability, gesture, facial expression and distance. The aim of the experiment is to demonstrate the capability of the system to drive the FACE robot. In this test the Kinect was placed above the robot head where the reference points of the Kinect sensor and robot are coincided with a little offset. The 2D and 3D images was captured in 30 fps in VGA resolution.

In general the equation of attention score can be expressed as below:

\[
Attention = K_1 F_1 \left( \frac{1}{\text{distance}} \right) + K_2 F_2 + K_3 F_3 \left( \frac{1}{\Delta t} \right) + K_4 F_4(\text{gesture}) + K_5 F_5(\text{facial expression}) 
\]  (1)

- \( \Delta t \): difference angle between sound source and subject.
- \( [K_{11},..., K_5] \): effectiveness of each stimulus on attention
- \( F_i \) : human-inspired functions that mimicking various attention models.

The calculated score is stored in the scene-meta image for each subjects present in the scene and will be used by the FACE control library to decide in which direction the gaze has to be oriented. The scene meta-image is streamed to the FACE control library using YARP [3].

Future Work
Future works will focus on the study of more human inspired attention models to be tested and simulated with the attention system through the design of dedicated \( F_i \) and \( K_i \).

Fig. 2. The attention of the system is changing based on the subjects’ events over the time.