A Cloud Architecture for Emotion Recognition Based on the Appraisal Theory

Marco Demutti[1], Vincenzo D'Amato[1], Carmine Tommaso Recchiuto[1], Luca Oneto[1] and Antonio Sgorbissa[1]

[1]*DIBRIS, Università di Genova, Via all'Opera Pia 13, 16145, Genova, Italy*

Abstract

Designing robots with the ability to infer a person's emotional state represents one of the major challenges in social robotics. This work proposes a cloud system for online human emotion recognition in spontaneous human-robot verbal interaction, structured as a set of REST API endpoints. Based on the appraisal theory of emotion, the system acquires data about the person's expected appraisal of a given situation, depending on their needs and goals, and combines it with sensory data, such as facial expressions, angles of the head, and gaze of the person, and distance between the person and the robot. The whole set of data is used to infer the person's emotional state during the interaction through a Random Forest classifier, trained for binary classification (i.e., positive vs. negative emotions). Results confirmed that using both data sources improved performance in both the K-fold and the Leave One Person Out scenarios.

Keywords

Human-robot interaction, social robotics, cloud robotics, REST API, emotion recognition, appraisal theory

1. Introduction

Providing a natural, genuine, and effective human-robot interaction (HRI) represents one of the major and fascinating challenges in Social Robotics [1]. The most crucial skill that confers naturalism to interactions between humans is our ability to infer the emotional states of others based on non-verbal signals, such as facial expressions, voice, body posture, and movement [2]. This ability allows us to adjust our social behaviors and communication patterns to optimize the interaction. A social robot with the same ability would reliably adapt to changes in its partners' behavior, and earn their trust during the interaction.

9th Italian Workshop on Artificial Intelligence and Robotics (*AIRO 2022)*

s4389233@studenti.unige.it (M. Demutti); vincenzostefano.damato@edu.unige.it (V. D'Amato); carmine.recchiuto@dibris.unige.it (C. T. Recchiuto); luca.oneto@unige.it (L. Oneto); antonio.sgorbissa@unige.it (A. Sgorbissa)

https://www.researchgate.net/profile/Marco-Demutti (M. Demutti); https://www.researchgate.net/profile/Vincenzo-Damato (V. D'Amato); https://www.researchgate.net/profile/Carmine-Recchiuto (C. T. Recchiuto); https://www.researchgate.net/profile/Luca-Oneto (L. Oneto); https://www.researchgate.net/profile/Antonio-Sgorbissa (A. Sgorbissa)

0000-0003-4285-1412 (M. Demutti); 0000-0002-2492-7340 (V. D'Amato); 0000-0001-9550-3740 (C. T. Recchiuto); 0000-0002-8445-395X (L. Oneto); 0000-0001-7789-4311 (A. Sgorbissa)

The complexity and variability of emotions make emotion recognition a challenging task, especially when performed in a natural and spontaneous HRI context, where conditions may diverge from the controlled environment where most experiments are carried out [3].

The vast literature on emotion recognition covers (i) the problem of emotion classification, discussing how emotions should be represented, and (ii) the choice of the most informative non-verbal signals for the robot to acquire and interpret. In other words, the two main research topics establish the output and the input of the classification process, respectively.

Narrowing it down to social robotics, most previous studies performed emotion recognition by combining multiple sensory modalities, such as facial expression, body posture, and speech, using them with black-box models to predict an emotion from a list of possible labels [4, 5, 6].

This approach only considers emotional expressions, as people reveal them to the outside world, voluntarily or not [7]. However, emotional expressions do not necessarily reflect the person's emotional state [8], which can be expressed in different ways depending on several individual factors or even hidden. For this reason, these techniques lead to good classification performances in acted situations but are less suitable in actual HRI scenarios.

This work proposes a novel emotion recognition framework to assess the person's emotional state during a dyadic autonomous HRI.

2. Emotion Recognition Through Cognitive Appraisal

The proposed emotion recognition framework is based on an implementation of the appraisal theory of emotion [9]. According to the theory, emotions result from a two-dimension individual evaluation (the so-called appraisal) of a person's situation. In the primary appraisal, the person evaluates the situation in terms of the relevance to their needs and congruence with their goals. The secondary appraisal mainly concerns the person's possibilities to cope with the situation.

Instead of recognizing emotions in people, the appraisal theory of emotion has frequently been employed to address the dual issue of generating and expressing emotions in artificial agents rather than recognizing them in individuals [10]: Kismet [11] represents one of the most significant studies in the field. However, the fact that the theory arose to predict human emotions supports extending its principles to emotion recognition.

In a previous study [12], we developed and used the emotion recognition framework to collect data from participants who had a spontaneous and autonomous verbal interaction with the humanoid robot Pepper, programmed to elicit different emotions in various moments of the conversation. Figure 1 shows the experimental setup[1].

Throughout the interaction, we combined information about the person's appraisal with state-of-the-art sensory data. More specifically, we trained a Random Forest classifier using two sources of data:

- Sensory data, which consisted of the user's facial expressions, head and gaze angles, and the distance from the camera.
- Appraisal data, which encoded information about the person's needs and goals and how coherent they were with what the robot said and did. For example, appraisal data

[1] A video showing participants during the experiment can be found here: https://youtu.be/73ecZZWgG0k

Figure 1: Experimental setup.

considered when the person decided to change the topic of conversation (which may indicate that the topic was not suitable for them, thus conflicting with their needs and goals) or how well the robot was able to perform the activity that the person had requested.

Binary classification (i.e., positive vs. negative emotion) results showed that using both data sources led to a performance improvement compared to using sensory data only. For example, the balanced accuracy passed from $(64.85 \pm 2.30)\%$ to $(66.44 \pm 0.55)\%$ and from $(59.71 \pm 1.33)\%$ to $(62.43 \pm 0.66)\%$, respectively.

3. Cloud Architecture For Emotion Recognition

Given these preliminary results, the current work proposes a cloud architecture for the online implementation of the system. The overall framework results from integrating an Emotion Recognition service with the preexisting CAIR verbal interaction system [13]. The CAIR system can manage a knowledge-based autonomous interaction by accepting commands to execute actions and conversing with the person about various topics. Such integration is possible due to the client-server architecture and the use of services implemented as REST APIs [14], which grant flexibility, scalability, portability, and independence. In the same way, further new services may be easily added in the future. The system can be used by most devices with Internet connectivity, able to acquire an input through a microphone and provide an output through a screen or speaker, combined with a camera (and possibly other sensors) acquiring data from the environment. Figure 2 shows the overall architecture of the system.

3.1. Server

The server is composed of three web services, implemented using the Flask-RESTful framework on Python[2]: i) the Hub service, that handles the requests from the client, ii) the Dialogue service, that manages the interaction with the user, and iii) the Emotion Recognition service, that provides the user's emotional state during the interaction.

[2]Flask-RESTful framework: https://flask-restful.readthedocs.io/en/latest/

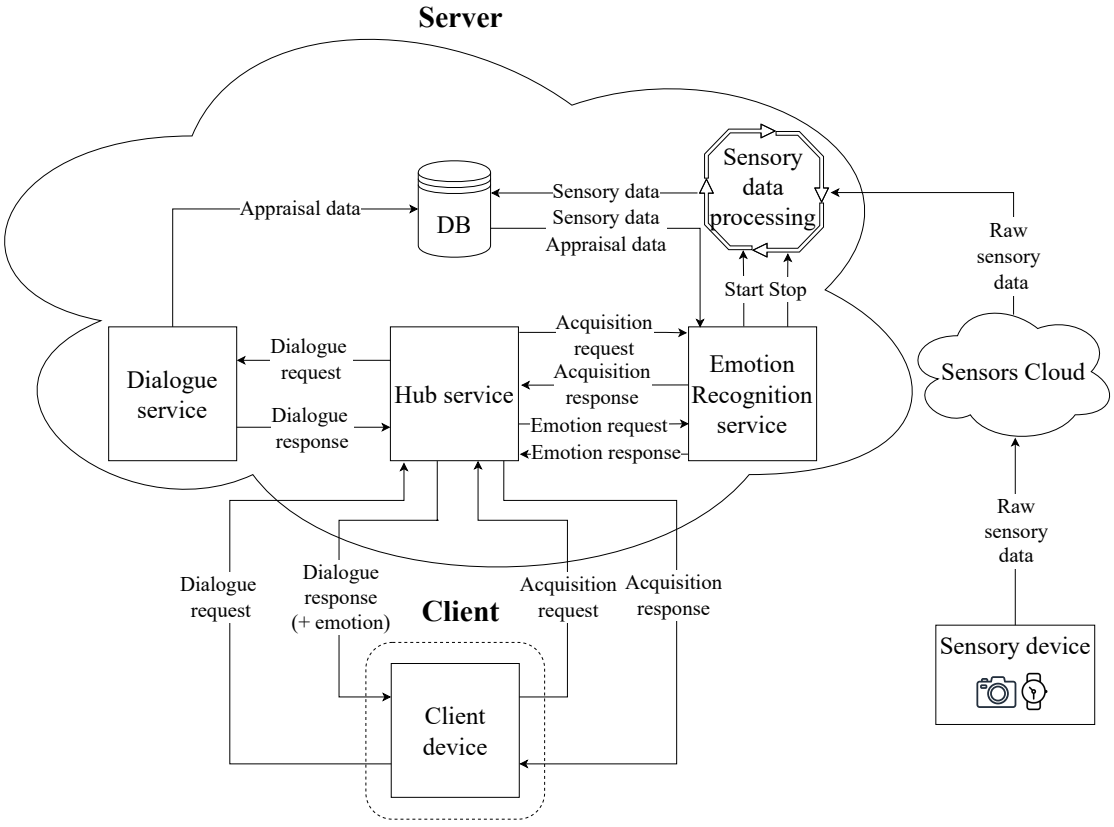

Figure 2: Emotion recognition framework.

3.1.1. Hub Service

The Hub service handles all the requests from the client. At the first request, the Hub associates the new client with an initial state, which will be stored on the client side and included in all upcoming requests. The client state is composed of the emotional state of the user and pieces of dialogue information, e.g., the current topic of conversation, the type of sentence chosen by the system, and the moments when the person and the robot started and finished speaking. Throughout the interaction, the requests from the client can be of two types: i) the "dialogue" request, aimed to develop the interaction with the user, and ii) the "acquisition" request to start or stop the acquisition from one or multiple sensors.

In case of a dialogue request, the Hub forwards it to the Dialogue service, which provides the next step of the interaction (more details in Section 3.1.2). Then, an "emotion" request, containing the client state, allows the Hub to obtain the user's emotional state from the Emotion Recognition service.

In case of an acquisition request, the Hub forwards it to the Emotion Recognition service (more details in Section 3.1.3), which starts or stops the acquisition from the corresponding sensor.

3.1.2. Dialogue Service

The Dialogue service is mainly responsible for managing the interaction with the user. It recognizes the person's intention to discuss a specific topic or to ask the agent to execute a task. More in detail, after processing the sentence pronounced by the person (contained in the dialogue request), it obtains the verbal reply and possibly the task to execute by exploiting the Ontology [13], containing all concepts and sentences used in the interaction. In addition, the service extract appraisal data from the user sentence and stores them in the SQLite database.

3.1.3. Emotion Recognition Service

The Emotion Recognition service exploits the pre-trained Random Forest classifier to assess the user's emotional state. It handles two types of requests from the Hub service, namely the acquisition and the emotion requests. Upon an acquisition request, the Emotion Recognition service starts or stops one or multiple "Sensory data processing" tasks, which process data coming from sensors through the cloud. Sensory data are continuously stored in the SQLite database during the acquisition. When the Hub sends an "emotion" request, the Emotion Recognition service retrieves the two categories of inputs of the classifier from the database, namely sensory and appraisal data (explained in Section 2). The emotion label is then returned to the Hub in the client state.

3.2. Client

As for the aim of this study, the client represents the robot used for the interaction. However, the client may also be a computer or, in general, most devices with Internet connectivity, able to acquire an input through a microphone and provide an output through a screen or speaker.

Each client is associated with a state, initialized at the first request to the Hub service, and then stored locally. The state is updated from the client side to contain helpful information, such as when the person and the robot started and finished speaking. The Dialogue and the Emotion Recognition services then use these pieces of information to provide the response to upcoming requests.

At the beginning of the interaction, the client also makes a request to the Hub service to start the acquisition from one or multiple sensors. For example, the request may include the IP address of the camera's video stream.

Throughout the interaction, once it has obtained the verbal reply and possibly the task to execute from the server, it interacts with the user and acquires their reply. The interaction ends when explicitly asked by the user.

3.3. Sensory Devices

Sensory devices acquire data and stream them to the cloud. Although the post-processing algorithm has been designed to extract data from camera video streams the system may be used for other types of sensors (such as a smartwatch or a microphone for speech emotion recognition).

References

[1] N. Mavridis, A review of verbal and non-verbal human-robot interactive communication, Robotics and Autonomous Systems 63 (2015) 22–35.

[2] M. Spezialetti, G. Placidi, S. Rossi, Emotion recognition for human-robot interaction: Recent advances and future perspectives, Frontiers in Robotics and AI 7 (2020).

[3] G. Castellano, I. Leite, A. Pereira, C. Martinho, A. Paiva, P. Mcowan, Affect recognition for interactive companions: Challenges and design in real world scenarios, Journal on Multimodal User Interfaces 3 (2009) 89–98.

[4] L. Chen, M. Zhou, W. Su, M. Wu, J. She, K. Hirota, Softmax regression based deep sparse autoencoder network for facial emotion recognition in human-robot interaction, Information Sciences 428 (2018) 49–61.

[5] M. Wu, W. Su, L. Chen, Z. Liu, W. Cao, K. Hirota, Weight-adapted convolution neural network for facial expression recognition in human-robot interaction, IEEE Transactions on Systems, Man, and Cybernetics: Systems 51 (2021) 1473–1484.

[6] C. C. Lee, E. Mower, C. Busso, S. Lee, S. Narayanan, Emotion recognition using a hierarchical binary decision tree approach, Speech Communication 53 (2011) 1162 – 1171. Sensing Emotion and Affect - Facing Realism in Speech Processing.

[7] M. Mortillaro, B. Meuleman, K. Scherer, Advocating a componential appraisal model to guide emotion recognition, International Journal of Synthetic Emotions 3 (2012).

[8] R. W. Picard, Affective Computing, MIT Press, 1997.

[9] R. S. Lazarus, Emotion and Adaptation, Oxford University Press, 1991.

[10] Z. Kowalczuk, M. Czubenko, T. Merta, Interpretation and modeling of emotions in the management of autonomous robots using a control paradigm based on a scheduling variable, Engineering Applications of Artificial Intelligence 91 (2020) 103562.

[11] C. Breazeal, Emotion and sociable humanoid robots, International Journal on Human-Computer Studies 59 (2003) 119–155.

[12] M. Demutti, V. D'Amato, C. Recchiuto, L. Oneto, A. Sgorbissa, Assessing emotions in human-robot interaction based on the appraisal theory, in: 2022 31st IEEE International Conference on Robot and Human Interactive Communication (RO-MAN), 2022, pp. 1435–1442.

[13] L. Grassi, C. T. Recchiuto, A. Sgorbissa, Sustainable verbal and non-verbal human-robot interaction through cloud services, 2022.

[14] M. Masse, REST API Design Rulebook: Designing Consistent RESTful Web Service Interfaces, O'Reilly Media, Inc., 2011.