A Preliminary Framework for a Social Robot "Sixth Sense"

Lorenzo Cominelli[1(✉)], Daniele Mazzei[1], Nicola Carbonaro[1], Roberto Garofalo[1], Abolfazl Zaraki[1], Alessandro Tognetti[1,2], and Danilo De Rossi[1,2]

[1] Faculty of Engineering, Research Center "E. Piaggio",
University of Pisa, Pisa, Italy
lorenzo.cominelli@for.unipi.it

[2] Department of Information Engineering, University of Pisa, Pisa, Italy
http://www.faceteam.it

Abstract. Building a social robot that is able to interact naturally with people is a challenging task that becomes even more ambitious if the robots' interlocutors are children involved in crowded scenarios like a classroom or a museum. In such scenarios, the main concern is enabling the robot to track the subjects' social and affective state modulating its behaviour on the basis of the engagement and the emotional state of its interlocutors. To reach this goal, the robot needs to gather visual and auditory data, but also to acquire physiological signals, which are fundamental for understating the interlocutors' psycho-physiological state. Following this purpose, several Human-Robot Interaction (HRI) frameworks have been proposed in the last years, although most of them have been based on the use of wearable sensors. However, wearable equipments are not the best technology for acquisition in crowded multi-party environments for obvious reasons (e.g., all the subjects should be prepared before the experiment by wearing the acquisition devices). Furthermore, wearable sensors, also if designed to be minimally intrusive, add an extra factor to the HRI scenarios, introducing a bias in the measurements due to psychological stress. In order to overcome this limitations, in this work, we present an unobtrusive method to acquire both visual and physiological signals from multiple subjects involved in HRI. The system is able to integrate acquired data and associate them with unique subjects' IDs. The implemented system has been tested with the FACE humanoid in order to assess integrated devices and algorithms technical features. Preliminary tests demonstrated that the developed system can be used for extending the FACE perception capabilities giving it a sort of sixth sense that will improve the robot empathic and behavioural capabilities.

Keywords: Affective computing · Behaviour monitoring · Human-Robot Interaction · Social robotics · Synthetic tutor

N.F. Lepora et al. (Eds.): Living Machines 2016, LNAI 9793, pp. 58–70, 2016.
DOI: 10.1007/978-3-319-42417-0_6

1 Introduction

Nowadays, it is well-known that our emotional state influences our life and decisions [1]. Education and learning processes are not excluded from this claim and the review, written by Bower [2], about how emotions might influence learning, as well as the more specific research about the effects of affect on foreign language learning, done by Scovel [3], are two of several important studies confirming this theory. As a consequence, in the last years this topic has triggered also the interest of social robotics scientists [4,5]. Indeed, they develop humanoid robots destined to interact with people, and the emotional and psychological states of these androids' interlocutors have to be necessarily taken into account. This becomes even more important in case where robots have to interpret the role of synthetic tutors intended for teaching children and conveying pedagogical contents in scenarios like musea or schools [6].

In order to interpret the emotional states of the pupils who interact with robots, social robotics researchers have proposed many different solutions that can be divided in two main categories: *Visual-Auditory Acquisition* and *Physiological Signals Acquisition*. The processing of both kinds of information aims to an estimation of the interlocutors' affective states, and these social/emotional information is exploited, in turn, by the control architecture of the robots, to modulate their behaviour, or completely change the actions they were planning. This improves the empathy and facilitates the dialogue between robots who are teaching and children who are learning. In any case, the main problem of these two approaches is that, up to date, there is not a good integration of data delivered by both the acquisitions. In most of the cases, the perception systems are designed to use only one kind of acquisition, and this is not sufficient to determine a stable perception of the human emotional state, but just a temporary assessment that can't be used in a long-term interaction [7].

The visual-auditory acquisition has the peculiarity to be more stable and well functioning thanks to a lot of available devices, which are easy to use but also easy to mislead (e.g., a smiling face has not to be always interpreted as an happy person); while the acquisition of physiological parameters has the advantage to reveal hidden emotional states and involuntary reactions that are relevant for human behaviour understanding, but this approach has the major outstanding problem to be an obtrusive, if not even invasive method.

In this paper, we present a novel architecture that supports the acquisition of both visual-auditory and physiological data. It is composed of the Scene Analyzer, a software for audio and video acquisition and social features extraction [8]; and the TouchMePad, consisting of two electronic patches and a dedicated signal processing software for the users physiological monitoring and affective state inference. TouchMePad is thought to be integrated with the other half of the perception system to become the sixth sense of a social robot. Moreover, the dedicated software, as well as the whole acquisition framework, has been designed to collect data in a sporadic and unobtrusive way in order to minimise the stress for the subjects and reach a high level of naturalness during the HRI. This is highly beneficial considering that Scene Analyzer is able to identify different persons by means of a QR codes reader module. Thanks to this recognition capability, once a subject is recognised, the physiological signals acquired by the TouchMePad can be stored in a database that will permanently associate a unique ID with the recognised subject. This information is compared with other meaningful social information (e.g., facial expression, posture, voice direction), providing the robot with the capability to change or modulate the tutoring according to different persons and their emotional state both in real time and accordingly to past interactions (e.g., previous lessons).

2 The Perception Framework

The perception framework is composed of two main parts: the Scene Analyzer, our open-source visual-auditory acquisition system that we continuously upgrade and deliver on Github[1]; the TouchMePad, composed of two hardware patches for physiological parameters acquisition and a dedicated software for signal processing and affective computing.

2.1 Scene Analyzer

We designed Scene Analyzer (SA) as an out-of-the-box human-inspired perception system that enables robots to perceive a wide range of social features with the awareness of their real-world contents. SA enjoys of a peculiar compatibility, indeed, due to its modular structure, it can be easily reconfigured and adapted to different robotics frameworks by adding/removing its perceptual modules. Therefore, it can be easily integrated to any robot irrespective of the working operating system.

SA consists of four distinct layers (shown in Fig. 1) data acquisition, low-level and high-level features extraction, structured data creation and communication. SA collects raw visual-auditory information and data about environment and, through its parallel perceptual sub-modules, extracts higher level information that are relevant from a social point of view (e.g., subject detection and tracking, facial expression analysis, age and gender estimation, speaking probability, body postures and gestures). All this information is stored in a dynamic storage called meta-scene. The SA data communication layer streams out the created meta-scene through YARP (Yet Another Robot Platform) middleware [9].

SA, supports a reliable, fast, and robust perception-data delivering engine specifically designed for the control of humanoid synthetic tutors. It collects visual-auditory data through Kinect ONE 2D camera, depth IR sensor, and microphone array with the highest level of precision. For example, it detects and keeps track of 25 body-joints of six subjects at the same time (Fig. 2), which is a fundamental capability for crowded scenarios. With such an accuracy it is possible to perceive gestures that are very important in educational contexts (e.g., head rubbing, arm crossed, exulting etc.). For the same purpose, extraction

[1] https://github.com/FACE-Team/SceneAnalyzer-2.0.

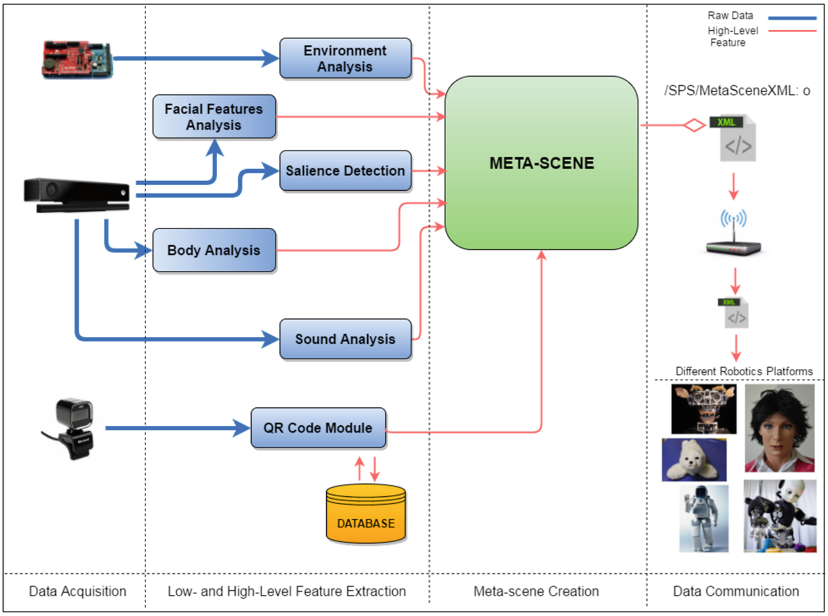

Fig. 1. Framework of the Scene Analyzer, the visual-auditory part of the perception system.

of data about joint coordinates of the lower part of the body has been added to the coordinates of the upper torso. These added coordinates makes the system capable to detect also the body pose of a person. Accordingly, we can distinguish between a seated and a standing position. For further information about meta-scene data structure and SA visual-auditory acquisition please refer to [10].

In order to endow the robot with a stable and reliable ID recognition capability, a new module has been implemented: the *QR Code Module*. It works in parallel analysing the image provided by a separated HD camera, and it is conceived for detecting and extracting information from QR codes. These bar-codes matrices can be printed and applied to the user's clothes or attached to the objects with which the robot has to interact. SA exploits an SQL database in which a list of QR codes are associated with permanent IDs and names. However, any further information can be added to any stored subject or object. Every time the synthetic agent will be able to recognise in the FOV a QR code that is known, because saved in the internal database, it will retrieve this information and automatically assign a known ID and name to that entity. Moreover, assuming that the entity is a person who has already interacted with the robot, once the subject has been recognised by the means of the QR code, the association between the permanent ID and that subject will continue even if the QR code will be no more visible by the camera.

This solution for assigning, storing and retrieving permanent information about interlocutors is fully automatic and mandatory in order to customise

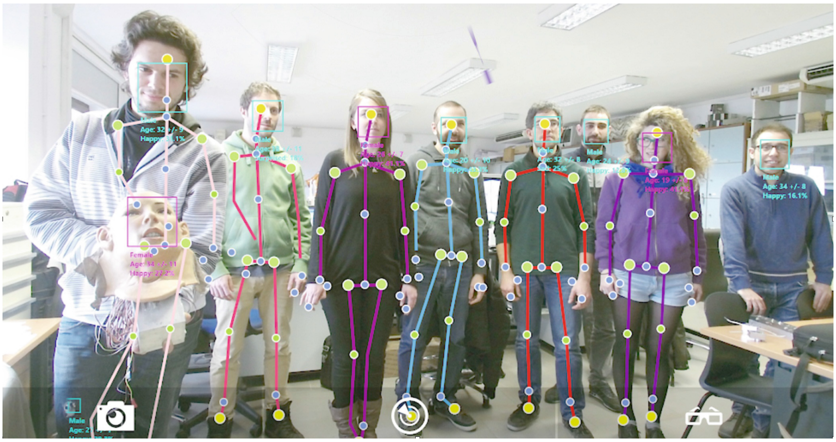

Fig. 2. Performance of the Scene Analyzer in a crowded scenario.

teaching methods for different students. The possibility to deliver personalised learning has been considered as a requirement according to the last decades trend in education and the perspectives for the next generation learning environment [11,12]. Another important feature of SA is processing of environmental data streamed by Zerynth Shield[2], an electronic board for environmental analysis (e.g., light intensity, temperature, audio level), which is useful to determine potential influences of the environment on interaction and learning engagement.

2.2 TouchMePad

In order to estimate and recognise human emotions from physiological parameters, several techniques have been developed in the last years, and most of them exploit wearable sensors (e.g., [13]). Since our system is intended to be used in crowded environments involving pupils or young users, the usage of sensorized clothes such as gloves or shirts is a considerable complication. Furthermore, a continuous and permanent contact would invalidate the naturalness of the interaction which is already difficult enough since a humanoid robot is implicated in it. Last but not least, an unceasing acquisition of multiple data from many subjects, including who is not currently involved in a real interaction with the robot, would be useless as well as overwhelming for the data processing phase.

For all these reasons, we opted for a user-centred solution that is non-invasive, unobtrusive and keeps the naturalness of a social interaction. Indeed, it is conceived to prevent discomfort for the user who has not to be permanently attached to sensors. On the contrary, two electronic patches are attached to the synthetic tutor shoulders and the subjects are asked to touch sporadically the shoulders of the robot in order to acquire physiological signals only in some key moments of

[2] http://www.zerynth.com/zerynth-shield/.

the interaction. This facilitates both the user and the acquisition system, reducing the number of contacts with the sensors, as well as the amount of gathered data, to the strictly necessary.

Therefore, TouchMePad (TMP) can be considered as the other half of the perception framework providing the robot with a sort of sixth sense. TMP is conceived to monitor the variation of the physiological parameters that are correlated to human affective state. As shown in Fig. 3 the system is composed of the electronic patches, acting as electrodes, and a central electronic unit for power supply, elaboration and transmission of user physiological parameters. The electronic unit is designed to conveniently combine ECG analog front-end with EDA one by means of a low-power micro controller. The developed ECG block is a three leads ECG system that samples signals at the frequency of 512 Hz. The front-end is based on the INA321 instrumentation amplifier and on one of the three integrated operational amplifiers available in the micro controller (MSP family made by Texas Instruments, MSP430FG439), to reach the total 1000x amplification. Regarding the EDA circuit, a small continuous voltage (0.5 V) is applied to the skin and the induced current is measured through two electrodes positioned in correspondence with the middle and the ring finger. The ratio between voltage drop and induced current represents the skin electric impedance, the EDA signal. An ad-hoc finger electrodes patches were designed and developed allowing the acquisition of user physiological parameters in a natural way. Therefore, the user does not have to wear any type of band, shirt etc. but simply touching the patches with the fingers it is possible to calculate the *Inter-Beat Interval* (IBI) parameter and the *Electro Dermal Activity* (EDA) signal. Finally, the system evaluates the robustness of user contact to identify physiological signal segments that are affected by artifacts and have to be discarded in further analysis. The detection of an artifact can also trigger an automatic request, by the robot to the user, to modify fingers position on the electrodes or to place them on the patches once again for a better acquisition.

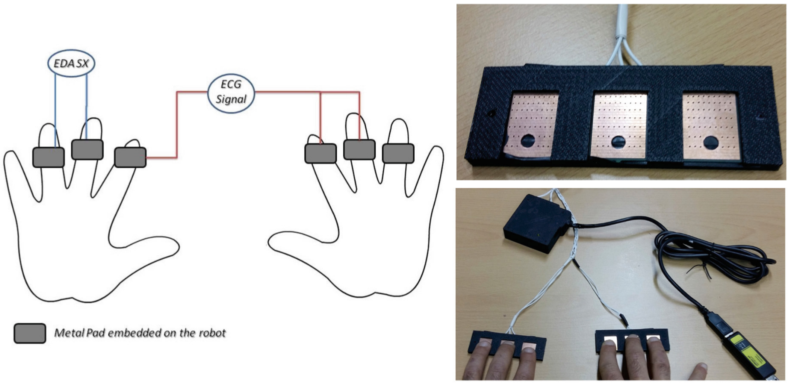

Fig. 3. Schema and photos of the acquisition patches.

For further information about physiologic signal acquisition and the methods we use for processing this data, please refer also to our previous works involving the acquisition of ECG [14,15] and EDA [16].

3 The ID Tracking Module

Since the perception system is composed of two separated main acquisition systems working in parallel (i.e., SA for visual-auditory acquisition and TMP for physiological parameters acquisition), we needed to find a solution to unify the data extracted from subjects in order to have a unique classification associated with IDs. The issue about having a link between information delivered by the SA and the one delivered by TMP depended on a very practical problem: every time a subject had to place their fingers on the electrodes, the camera got covered by that subject, and the SA was no longer able to detect any person in the field of view. This entails that the entire perception system, as it was initially conceived, was not able to assign the physiologic parameters to any person, because no person was detected while signals were acquired. Therefore, we developed a dedicated sub-module, which provided a workaround for this ‘identity problem’, ensuring the assignment of both SA and TMP data to a univocal ID.

We named it the *ID Tracking Module* (IDTm) and it works as a bridge between the TMP and the SA. This module continuously calculates the standard deviation of the pixels extracted by a cropped, re-sized, central part of the image acquired by the Kinect camera. Every time a subject approaches the patches covers most of the image field, as well as all the other people present in the scene. Nevertheless, we have another effect: the image becomes almost the same, especially the central part, regardless light conditions, contrast and brightness. Furthermore, the SA delivers information about the distance in meters of any subject until is detected. Considering that only one subject at a time has the possibility to put their fingers on the patches, we retrieve the information about the last detected closest subject and, when the standard deviation of the central image gathered by the camera comes under a threshold (empirically decided at 50), the IDTm saves the ID of the last detected closest subject, assumes that he/she is the one approaching the pads, and will assign potential physiologic parameters detected by the TMP, to that specific ID. All the information about the physiological parameters is stored in the database, assigned to the $ID_{Touching}$ provided by the IDTm, ready for an affective state estimation and a possible comparison between past and future acquisitions. Finally, the subject is detected again and recognised as soon as he leaves the patches and comes back clear in the gathered image.

This strategy not only solves a system drawback, but also demonstrates the power of a multi-tasking and modular architecture as the one we are presenting. A schema of this solution is shown in Fig. 4.

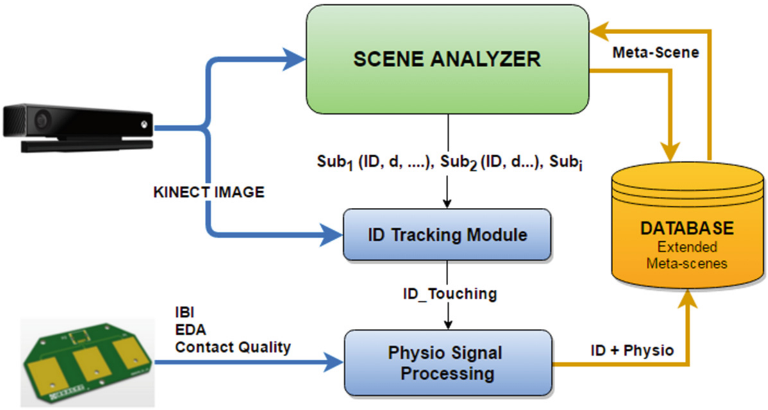

Fig. 4. The ID tracking schema.

4 Evaluation of the Perception System

4.1 Materials and Methods

The use case is based on the installation of the SA and the TMP on the FACE Robot body [17].

Data collected by the SA are stored in an SQL database together with the one provided by the TMP, thanks to the ID Tracking module. Once a contact on the TMP is detected, the quality of the signal is analysed and the information about contact quality is shown on the screen in which the interlocutor can read instructions. Projected instructions are: *"Please put your fingers on the patches"*, *"Bad contact! Please place your fingers again"* or *"Good Contact! Please keep your fingers on the patches"*. In this latter case, a 60 s countdown immediately starts on the screen. This is the duration that we set for having a reliable acquisition of physiological data. As a consequence, if the contact is maintained at a sufficient quality level, at the end of the countdown it does appear the following message: *"Thank you, you can remove your hands now"*. Acquired data are pre-filtered by the *Physio Signal Processing Module* shown in Fig. 4, which discards unreliable values, then calculates the IBI mean value, the EDA, asks for the $ID_{Touching}$ to the IDTm, then finally sends through yarp all this information to the database, adding the physio data to the proper ID.

Several tests were performed in order to verify the capability of the TMP to monitor user physiological state. A 27 years old male was selected for the experiment whose heart beat frequency in rest condition is known and in value of 80 bpm. This information has been useful for the comparison with the value calculated from the IBI values extracted by the sensor patches. The test consisted of different sessions of 1 min in which the subject was asked to maintain a rest condition and to put his fingers of both hands in direct contact with the TMP.

4.2 The Experiment

The experiment begins with the detection of a social scene involving two subjects. Initially, the image gathered by the SA has a high variance as shown in the first chart of Fig. 5. After 5 s the robot is able to detect both subject 1, who is a 26 years old female (the red line in Fig. 5), and subject 2 (the blue line in the same chart), which is the 27 years old male who is going to touch the TMP. After 8 s subject 2 approaches the FACE Robot and his detected distance decreases, while the female remains standing at 1.6 m from the robot. When subject 1 gets closer to the robot there are two consequences: covering the image, the pixels' standard deviation decreases and, at 10 s, subject 1 gets lost by the perception system. After just 1 s, also subject 2 disappears completely from the scene ($d = 0$), but his ID is saved by the IDTm as the last closest subject's ID, assuming that he is the one who is going to touch the patches.

Therefore, the acquisition system starts looking for a potential contact, checking and analysing the contact quality of the finger electrodes. This can be noticed in the first 2 s of Fig. 6, that shows the trend of the *Contact Quality* (CQ) parameter in correspondence with the two acquired physiological parameters (i.e., IBI and EDA). CQ allows to distinguish not only between bad and good sensor contact, but also to determine which fingers are not well positioned by means of the elaboration of the relative physiological signal. In fact, each of the six fingers involved in the acquisition has some peculiarities: for example the right hand medium finger is used as the reference electrode for the ECG circuit, while the ring and the medium finger of the left hand allow to track the EDA signal, as previously shown in the schema of Fig. 3. Considering Table 1, in which all the possible combination of CQ values are reported, we can claim that the system is working correctly when the CQ parameter is equal to 50, while in other condition the validity of physiological data is not guaranteed and fingers should be repositioned.

The physiological parameters gathered during all the 60 s of acquisition are shown in Fig. 6. At the beginning CQ is 60, representing that the circuit for hand detection is activated and ready for the elaboration. Then the IBI and EDA values are tracked and stored. At $t = 9s$ the CQ value decreases to 30,

Table 1. CQ value definition

CQ value	Meaning
0	*No Contact*
10	*IBI No Contact, EDA Good Contact*
20	*IBI & EDA Bad Contact*
30	*IBI Bad Contact, EDA Good Contact*
40	*IBI Good Contact, EDA Bad Contact*
50	*IBI & EDA Good Contact*
60	*Contact Circuit Active*

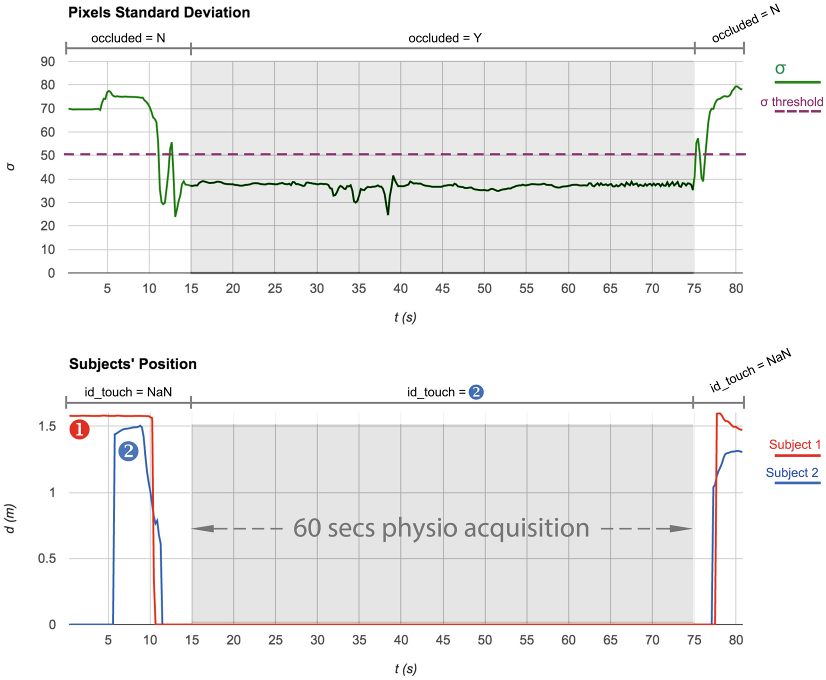

Fig. 5. The social robot sixth sense evaluation test - ID tracking module. (Color figure online)

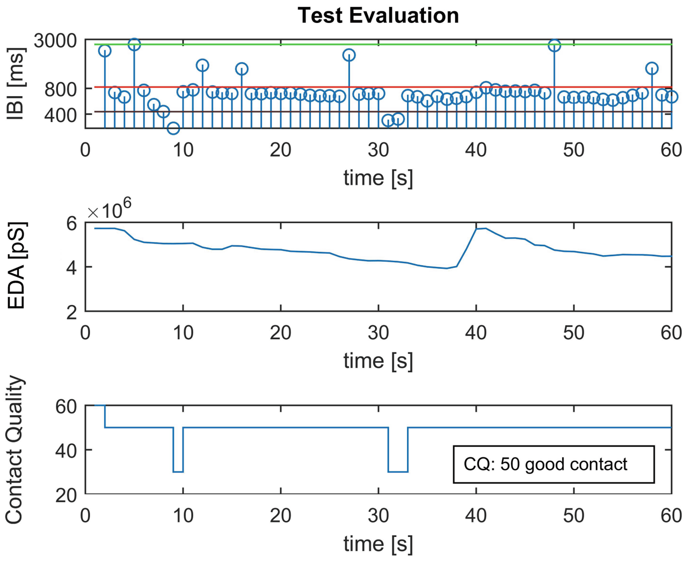

Fig. 6. Physiological Parameters extracted by the perception framework in real time during the 60 s of acquisition highlighted in Fig. 5.

likely due to a bad placement or a movement of the finger for IBI acquisition. At the same instant, in fact, it is possible to notice that the IBI signal presents a sudden variation with its value that goes under the minimum level of the session (represented by the green line). The same situation could be found at $t = 31s$. An other important event is located at $t = 40s$. At this stage the subject was elicited with an auditory stimulus (i.e., a sudden hand clap) to validate the contribution of the EDA circuit. As a results, in Fig. 6 the EDA signal leaves the constant trend with a relative peak that depends on the phasic activity of the sweat gland in response to the auditory stimulus. At the end of one minute of acquisition the subject is asked to remove his hands from the patches. He moves away from the robot that detects again him and the female who was still standing in the initial position, recognising both of them.

5 Conclusions and Future Works

In this preliminary work, an unobtrusive social scene perception system provided with a physiological signal perception “sixth sense” given by a sporadic acquisition of physiological parameters, has been developed. Such a system has been proved to endow a social robot with the possibility to infer emotional and psychological states of its interlocutors. This information is fundamental for social robots aimed at establishing natural and emphatic interactions with humans. This system has been designed in order to be used in robotic enhanced teaching contexts, where social robots assume the role of synthetic tutors for children and pupils. To evaluate the developed perception framework including all its features, we designed an ad-hoc test that stressed all the functionalities of the acquisition system. The experiment demonstrated the capability of the system to properly acquire the entire data-set of information assigning gathered data to unique and permanent subjects’ IDs. Moreover, it is important to highlight that the fusion of the data collected by the two system integrated in the presented setup goes beyond the simple data merging. The enhanced meta-scene which results from the system will give to researchers the possibility to better infer the subject’s social and affective state by correlating psycho-physiological signals with behavioural data.

The presented perception system will be used as the acquisition system of several humanoid robots involved in the EASEL European Project[3] in different educational scenarios (e.g., musea, school classrooms). This will validate the portability of the system and will make it a practical framework for testing how to improve the dialogue between robotic tutors and children as well as their learning.

Acknowledgment. This work was partially funded by the European Commission under the 7th Framework Program projects EASEL, Expressive Agents for Symbiotic Education and Learning, under Grant 611971-FP7- ICT-2013-10. Special thanks to Daniela Gasperini for her fundamental contribution in the experiments organization.

[3] http://easel.upf.edu/.

References

1. Damasio, A.: Descartes' Error: Emotion, Reason, and the Human Brain. Grosset/Putnam, New York (1994)
2. Bower, G.H.: How might emotions affect learning. Handb. Emot. Mem. Res. Theor. **3**, 31 (1992)
3. Scovel, T.: The effect of affect on foreign language learning: a review of the anxiety research. Lang. Learn. **28**(1), 129–142 (1978)
4. Hudlicka, E.: To feel or not to feel: the role of affect in human-computer interaction. Int. J. Hum. Comput. Stud. **59**(1), 1–32 (2003)
5. Fong, T., Nourbakhsh, I., Dautenhahn, K.: A survey of socially interactive robots. Robot. Auton. Syst. **42**(3), 143–166 (2003)
6. Causo, A., Vo, G.T., Chen, I.M., Yeo, S.H.: Design of robots used as education companion and tutor. In: Zeghloul, S., Laribi, M.A., Gazeau, J.-P. (eds.) Robotics and Mechatronics. Mechanisms and Machine Science, vol. 37, pp. 75–84. Springer, Switzerland (2016)
7. Yan, H., Ang Jr., M.H., Poo, A.N.: A survey on perception methods for human-robot interaction in social robots. Int. J. Soc. Robot. **6**(1), 85–119 (2014)
8. Zaraki, A., Mazzei, D., Giuliani, M., De Rossi, D.: Designing and evaluating a social gaze-control system for a humanoid robot. IEEE Trans. Hum. Mach. Syst. **44**(2), 157–168 (2014)
9. Metta, G., Fitzpatrick, P., Natale, L.: Yarp: yet another robot platform. Int. J. Adv. Robot. Syst. **3**(1), 43–48 (2006)
10. Zaraki, A., Giuliani, M., Dehkordi, M.B., Mazzei, D., D'ursi, A., De Rossi, D.: An rgb-d based social behavior interpretation system for a humanoid social robot. In: 2014 Second RSI/ISM International Conference on Robotics and Mechatronics (ICRoM), pp. 185–190. IEEE (2014)
11. Sampson, D., Karagiannidis, C.: Personalised learning: educational, technological and standardisation perspective. Interact. Educ. Multimedia (4) 24–39 (2010)
12. Brusilovsky, P.: Developing adaptive educational hypermedia systems: from design models to authoring tools. In: Murray, T., Blessing, S.B., Ainsworth, S. (eds.) Authoring Tools for Advanced Technology Learning Environments, pp. 377–409. Springer, Netherlands (2003)
13. Lisetti, C.L., Nasoz, F.: Using noninvasive wearable computers to recognize human emotions from physiological signals. EURASIP J. Adv. Sign. Process. **2004**(11), 1–16 (2004)
14. Tartarisco, G., Carbonaro, N., Tonacci, A., Bernava, G., Arnao, A., Crifaci, G., Cipresso, P., Riva, G., Gaggioli, A., De Rossi, D., et al.: Neuro-fuzzy physiological computing to assess stress levels in virtual reality therapy. Interact. Comput. (2015). iwv010
15. Carbonaro, N., Anania, G., Mura, G.D., Tesconi, M., Tognetti, A., Zupone, G., De Rossi, D.: Wearable biomonitoring system for stress management: a preliminary study on robust ECG signal processing. In: 2011 IEEE International Symposium on a World of Wireless, Mobile and Multimedia Networks (WoWMoM), pp. 1–6. IEEE (2011)

16. Carbonaro, N., Greco, A., Anania, G., Dalle Mura, G., Tognetti, A., Scilingo, E., De Rossi, D., Lanata, A.: Unobtrusive physiological and gesture wearable acquisition system: a preliminary study on behavioral and emotional correlations. In: Global Health, pp. 88–92 (2012)
17. Mazzei, D., Zaraki, A., Lazzeri, N., De Rossi, D.: Recognition and expression of emotions by a symbiotic android head. In: 2014 14th IEEE-RAS International Conference on Humanoid Robots (Humanoids), pp. 134–139. IEEE (2014)