future internet

MDPI

Article

A Multimodal Perception Framework for Users Emotional State Assessment in Social Robotics

Lorenzo Cominelli [1,*], Nicola Carbonaro [1,2], Daniele Mazzei [1,3], Roberto Garofalo [1], Alessandro Tognetti [1,2] and Danilo De Rossi [1,2]

[1] Research Center "E. Piaggio", Faculty of Engineering, University of Pisa, 56122 Pisa, Italy; nicola.carbonaro@centropiaggio.unipi.it (N.C.); daniele.mazzei@centropiaggio.unipi.it (D.M.); roberto.garofalo@centropiaggio.unipi.it (R.G.); a.tognetti@centropiaggio.unipi.it (A.T.); d.derossi@centropiaggio.unipi.it (D.D.R.)

[2] Department of Information Engineering, University of Pisa, 56122 Pisa, Italy

[3] Department of Computer Science, University of Pisa, 56127 Pisa, Italy

* Correspondence: lorenzo.cominelli@centropiaggio.unipi.it; Tel: +39-050-221-7050

Received: 20 June 2017; Accepted: 26 July 2017; Published: 1 August 2017

Abstract: In this work, we present an unobtrusive and non-invasive perception framework based on the synergy between two main acquisition systems: the Touch-Me Pad, consisting of two electronic patches for physiological signal extraction and processing; and the Scene Analyzer, a visual-auditory perception system specifically designed for the detection of social and emotional cues. It will be explained how the information extracted by this specific kind of framework is particularly suitable for social robotics applications and how the system has been conceived in order to be used in human-robot interaction scenarios.

Keywords: perception framework; physiological signal acquisition; multimodality; social robotics; human-robot interaction

1. Introduction

Today, assessing the emotional state of a user can be considered one of the most important and challenging aims in human-computer interaction (HCI). In particular, this became a hot topic from the moment at which the influence of our emotional state in our life and decisions has been scientifically explained and demonstrated [1]. As a consequence, many computer scientists and researchers around the world are engaged in the development of machines that recognise, express, model, communicate, and respond to emotional information. This kind of information serves to let the systems adapt in real time their interaction depending on specific states or changes of state of the users. In the last decades, there has been a lot of progress in this direction, and several scientists claim that in the foreseeable future we will live and collaborate with machines that will be able to estimate our emotional state, ideally at the same level that people can [2,3]. On the other hand, it is persistent in the scientific community the skepticism about the practical implementation of such machines and their effective capability of sensing and recognizing emotions [4]. The main criticism is that the range of means and modalities of emotion expression is too broad, with many of these modalities being inaccessible (e.g., blood chemistry, brain activity, neurotransmitters), and many others being too non-differentiated [5]. Nonetheless, the simultaneous usage of multiple modalities for acquiring information about the users emotional state has proved to be a smart solution to the problem, especially when they involve computer vision and image processing [6]. Typically, these audiovisual acquisition systems are focused on face detection [7,8], face recognition [9], facial expression analysis [10,11], vocal emotion [12,13], gesture recognition [14,15], human motion analysis [16–18], audiovisual automatic speech recognition [19], and eye tracking [20,21]. Reviews of vision-based HCI are presented in [22], while a deepened study on

adaptive and intelligent HCI is discussed in [23] with a review of computer vision for human motion analysis, and a discussion of techniques for lower arm movement detection, face processing, and gaze analysis. In the recent years, the integration of these acquisition modalities led to the development of many multimodal interfaces, so much to become an increasingly important research direction, defined by Turk et al. as *Perceptual Interfaces* [24].

There are many different areas in which perceptual interfaces are implemented in order to estimate users emotional state. Biofeedbacks are exploited as an additional input in affective video gaming for modeling enjoying preference [25] or let the game adapt to the user emotional responses [26]. Ambient intelligence and domotics are often based on gesture recognition and physiological parameters extraction, not only as control inputs, but also for the user emotional state estimation [27]. Similarly, there are several important applications in affect recognition for computer-assisted intervention of children with Autism Spectrum Disorder (ASD) that are based on physiological signals extractions [28] or behaviour monitoring via face, gesture and posture recognition [29]. In any case, what emerges from the study of these more practical implementations is that, when it comes to real applications, having one highly reliable modality of acquisition is preferred than a true integration of many modalities, even if this paradigm is in contradiction with the previous mentioned problem: one single modality is definitely not sufficient for the estimation of a human emotional state. For instance, a peak in the electrodermal activity does not have a qualitative meaning, a specific gesture without knowing the correspondent facial expression can be ambiguous.

Moreover, in almost all the cited works, the instrumentation for physiological parameters acquisition, as well as for computer vision techniques, is very evident, if not obtrusive. Such acquisition frameworks are not usable in some specific cases. In particular, when moving from HCI to human-robot interaction (HRI). In this field, it is very important to have a non-invasive perceptual interface, in order to keep the interaction between the human interlocutor and the synthetic agent as natural as possible. Furthermore, whether the agent is a social robot, the interaction has not to be only natural but even elicited [30]. In social robotics, the understanding and the expressiveness of emotions are critical for the believability of the robots [31], particularly because of their applications: elderly care [32,33], ASD robot therapy [34–36], educational contexts [37]. In all these scenarios, the robot has to be highly adaptive and responsive to the psycho-physical state of the interlocutor. Peculiarly, in the educational context (e.g., schools, museums), in which the robot plays the role of a synthetic tutor. In such cases, it is not possible to rely on a structured environment, the users are many and diversified, and the capability to establish long-term interactions is needed [38].

Following all these requirements, here we present a novel multimodal perception framework specifically designed for social robots destined to interact with people in an non specifically predictable environment. The perception is based on two main acquisition modalities, audiovisual acquisition and physiological parameter extraction. The aim is to provide the robot with a structured data-set about the non-structured social environment in which is involved. In particular, in this work, we underline some fundamental features of the perception system: it is non-invasive, unobtrusive, designed to be embedded in the body of the robot, reliable but reduced to the minimum essential in order to be not consuming in terms of time and computational costs. Finally, we show two experiments: one that validates our proposed engineered solution for keeping track of the subjects, the other demonstrating the actual integration of the information gathered by both of the two main acquisition modalities.

2. Material & Methods

In order to realise the presented perception framework, we integrated the Scene Analyzer (SA; V2.0, FACEteam, Pisa, Italy), our software for audio and video acquisition and social features extraction, with the Touch-Me Pad (TMP), composed of two electronic patches able to extract physiological signals by fingers contact and its dedicated signal processing software for the user physiological monitoring. The TMP can be considered the added value that endows a social robot with a kind of sixth sense. The peculiarities of its design concept is the possibility to physically integrate the patches in the body of a

robot, and the way it collects data. Indeed, we are using a sporadic acquisition, minimising the stress and maximising the naturalness of the HRI. In this section we will present a brief description of the Scene Analyzer, which is deeply discussed in a previous publication [39], while we will examine in depth the Touch-Me Pad, in terms of design, development, technical description and validation, as well as the ID (Identifier) tracking strategy, that represents the bridge between SA and TMP and gives the system the possibility to be used for long-term interactions.

2.1. Scene Anayzer

We developed Scene Analyzer (SA) to be the audiovisual part of the perception system. It is human-inspired and specifically designed for social features extractions in natural interaction scenarios. SA can be considered the eyes and the ears of a social robot. It is based on the processing of raw visual-auditory information gathered by the microphone array and the RGB-D camera of a Microsoft Kinect ONE sensor, and a separated HD (High Definition) camera. The software is an integration of many sub-modules, each of which is dedicated to a specific analysis of the detected sound or image. SA is able to keep track of 25 joints of the body of 6 subjects at the same time in order to extract very important social cues, such as meaningful gestures and postures of the interlocutors (e.g., waving, exulting, head rubbing, crossing of the arms, etc.). The software provides also facial expression estimation, age and gender assessment—even for more than 6 persons - speaker identification, proximity, noise level and much more (Figure 1). Thanks to its modular structure, SA can be easily reconfigured and adapted to different robotic frameworks by adding or removing its perceptual modules.

As shown in Figure 2, it is possible to distinguish 4 layers describing the information flow of the SA from the physical sensors to a structured data-set: data acquisition, social features extraction, meta-scene creation and data communication, for which we use the YARP middleware [40]. It is important to emphasise that, regardless which are the sources gathering the raw data from the environment and the sub-modules processing this heterogeneous information, the output of the SA will be the unique specific file called *Meta-Scene*. This output is agnostic of both the original input sensors and the actuation system that are connected. This is fundamental to ensure the portability of the system. The meta-scene is processed and streamed every 300 ms, and it represents the organised knowledge of the robot about the noisy environment in which is involved. This data structure is divided by entities (i.e., *environment*, *objects* and *subjects*) and each entity has a unique ID and it is represented as a sub-structure of the meta-scene. For instance, a person in the meta-scene is a *subject*, with a specific ID, containing all the information that the perception system is able to perceive about that person in that precise moment. Since the SA extracts social meaning and behavioral data from the gathered raw data, we can assume that provides a social type of information, and SA can be defined as a social visual-auditory perception system. The meta-scene must be, not exclusively the output of SA, but the final result of the whole perception framework. In Section 2.3, we explain how we extend the meta-scene with the physiological data of the subjects, provided by the TMP, and how this continuously upgraded working memory is connected to a fixed storage that includes the data we want to keep for a longer period of time. This is how we realised an actual integration of the two acquisition systems and how we solved the problems of the identification and storage for long-term acquisitions.

A detailed description of the Scene Analyzer is available in our previous works [39,41], where it is also provided the validation of the software and a useful instrument for the arrangement of HRI experiments with SA. Moreover, SA is delivered as an open-source software available on Github (https://github.com/FACE-Team/SceneAnalyzer-2.0).

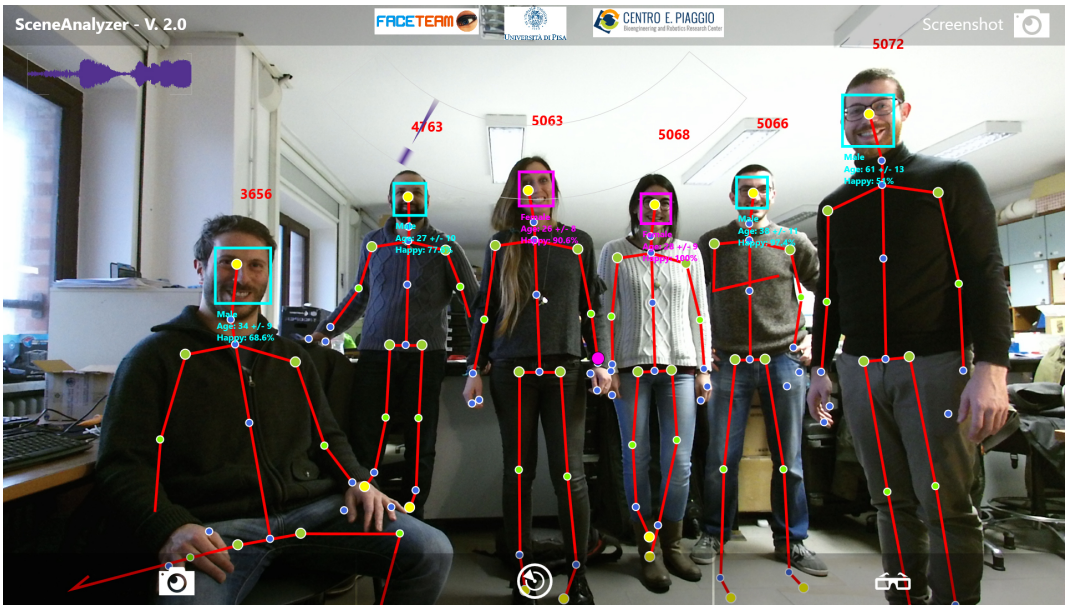

Figure 1. Screenshot of the SA (Scene Analyzer) visualiser in a case of crowded scenario.

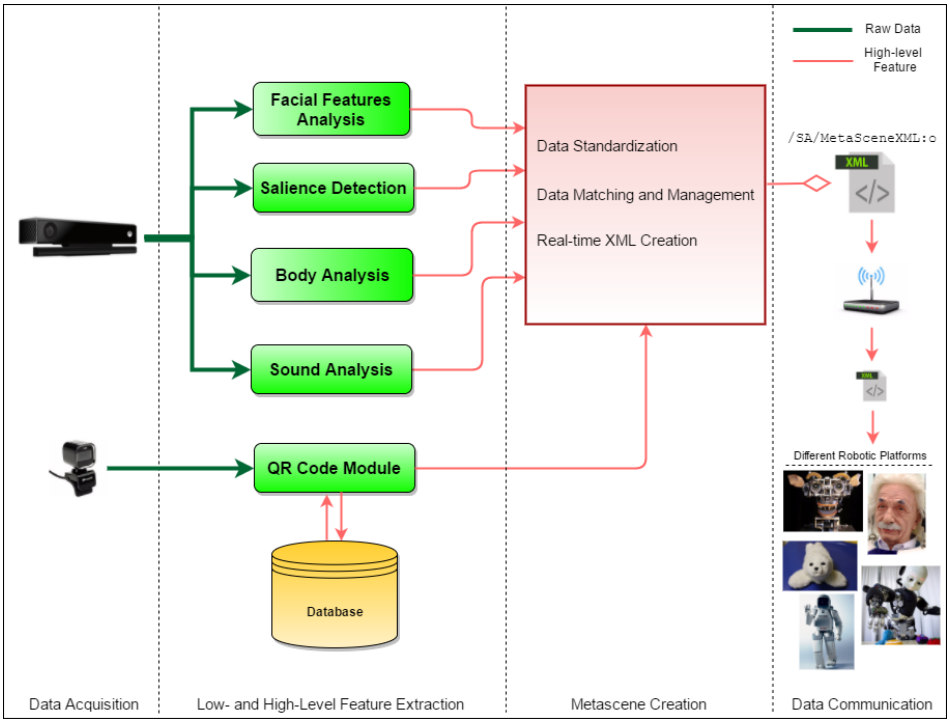

Figure 2. Sceme Analyzer framework and information flow; QR: Quick Response.

2.1.1. Touch-Me Pad

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., [42]). Since our system is intended to be used in crowded environments, involving subjects like pupils or young users, the usage of sensorised 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 robot will be 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 and 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. 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.

Touch-Me Pad (TMP) can be considered as the other half of the perception framework. It is conceived to monitor the variation of the physiological parameters that are correlated to human affective state. 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 design of the electronic system and of the electrode materials was conceived for making the acquisition set-up more convenient in the target scenario. An other important aspect taken into account during the patch developing phase was the idea to avoid the use of common gel electrodes for the acquisition of the physiological signals.

The electronic patches are composed by a thin and flexible PCB material (0.3 mm thick) in which large metal pads (15 mm $\times$ 20 mm) act as finger electrodes. The physiological signals selected were Electro Dermal Activity (EDA) and Inter-Beat Interval (IBI). The EDA signal have been widely used in literature for the assessment of human emotion [43–45], and classically is measured in the fingers site [46]. The IBI signal, representing the time in millisecond between two consecutive R waves of the electrocardiography (ECG), is an important parameter for the study of ECG correlates, like heart rate (HR) and heart rate variability (HRV), and is usually acquired with disposable Ag/AgCl electrodes placed in the torax area [47]. The idea was to develop an architecture in which the use of the user fingers for the simultaneous acquisition of EDA and IBI parameters could be combined (as shown in Figure 3).

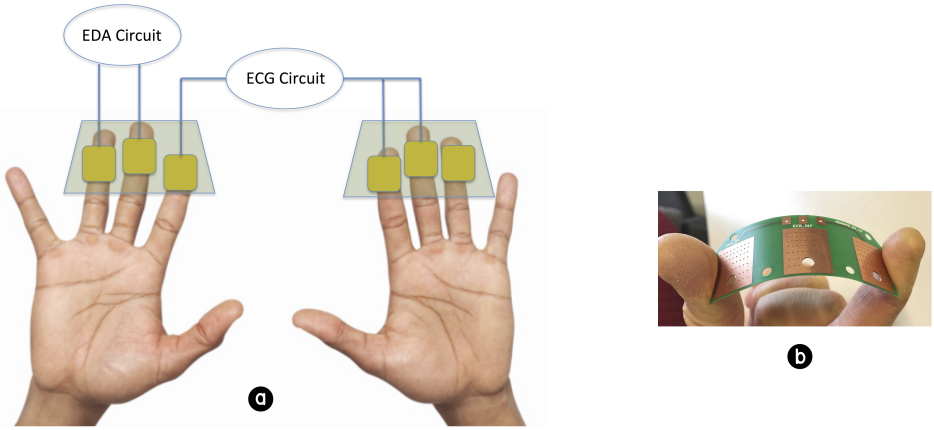

Figure 3. Schema of the acquisition patches. In (a) is represented the working principle of the finger electrodes, while in (b) is shown the developed patch and its flexibility characteristics; EDA: Electro Dermal Activity; ECG: electrocardiography.

This configuration allows to acquire user physiological parameters in a natural way. In fact, the user does not have to wear any type of band or shirt, but simply touching the patches with the fingers it is possible to calculate the IBI value and the EDA signal. In our prototype, we developed an electronic front-end to determine bad, good or no contact of the subject's finger tips with the conductive pads. The monitoring process begins by injecting a known DC signal into the ECG and EDA analog front-end and monitoring the signal changes when coming into contact with the input load. When contact is made, the current flows through the path causing a variation on the voltage of the conductive pad. This voltage is sampled by an analog-to-digital converter (ADC) channel

(embedded on the MSP430 microcontroller) and the digital conversion result is compared against a set of thresholds to determine good, bad, or no contact. This is the way 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. We named this parameter *Contact Quality* (CQ) (Table 1).

The detection of an artifact could also trigger an automatic request to modify fingers position on the electrodes or to place them on the patches once again for a better acquisition session.

Table 1. CQ (contact quality) Value Definition; IBI: Inter-Beat Interval; EDA: Electro Dermal Activity.

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

The electronic unit integrates the analog front-end required for ECG and EDA measurements, and a low-power micro controller for digital conversion and data elaboration. The ECG block is a three leads ECG system that samples signals at the frequency of 512 Hz. The analog 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 1000× amplification. The EDA measurement methodology is based on the evaluation of the variation of the skin electric impedance. The EDA analog front-end provides a small continuous voltage (0.5 V) to the skin and the induced current is measured through two electrodes positioned in correspondence with the middle and the ring finger. The relation between constant voltage applied and induced current represents the user skin electric impedance.

2.2. The Robotic Platform

Considering the relevance of physical embodiment and its measurable effect on performance and perception of social interactions [48,49], we decided to integrate all the sensors used by TMP and SA in a physical robotic platform: the higlhy expressive humanoid robot FACE [50,51]. FACE was initially a woman-like robotic head designed and developed by Hanson Robotics (http://www.hansonrobotics.com/), capable to convey emotion by means of the movement of a hyper-realistic facial expression controlled by 32 servomotors (Figure 4). These motors guide both the expression and the movement of a 4 dof neck.

On the basis of the fixed dimensions of the robotic head, we designed a body that respects anthropometrical proportions. The components have been downloaded and modified starting from parts of the InMoov project (http://inmoov.fr/), an open-source project that let researchers and makers develop a human-scaled 3D printable android. We customised the central part of the torso to create a suitable recess for the Kinect ONE sensor (Figure 5), we applied the Microsoft HD small camera on the top of the torso, and the TMP electronic patches have been integrated in the upper part of the shoulders, with specific housing for 3 finger contacts each side (Figure 6).

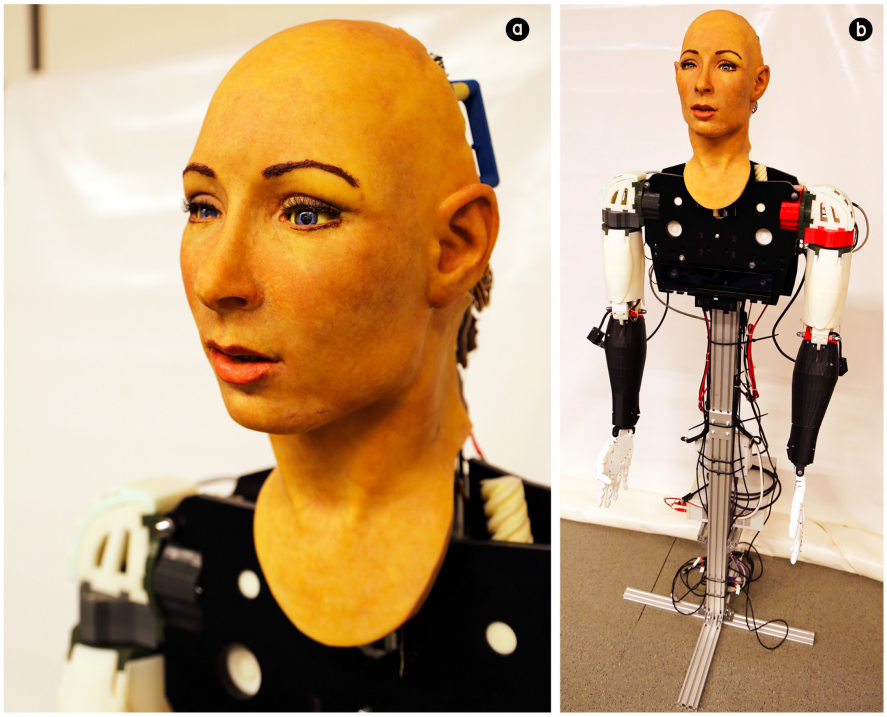

Figure 4. The FACE Robot and its novel body; (**a**) a close-up portrait; (**b**) a full-body portrait.

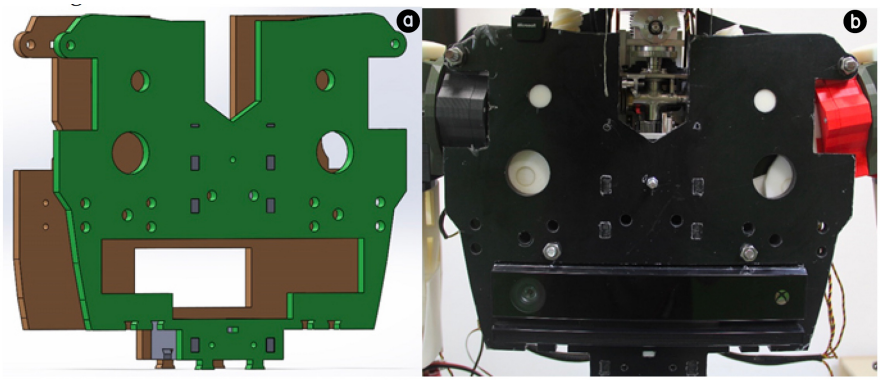

Figure 5. Development and integration of the SA sensors in the torso designed for FACE; (**a**) Computer Assisted Design (CAD) project of the FACE torso; (**b**) realisation of the torso.

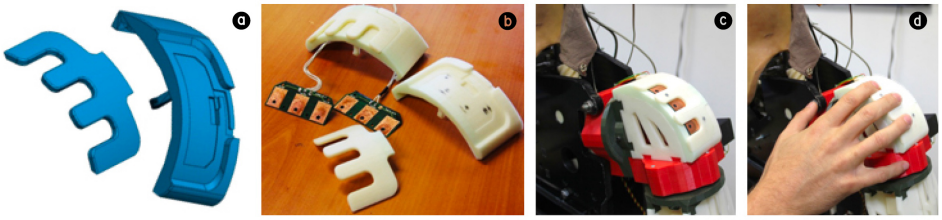

Figure 6. Development and integration of the TMP electronic patches in the shoulder of the body designed for FACE; (**a**) CAD project of the TMP patches; (**b**) realisation of the TMP patches; (**c**) integration of the TMP patches in the FACE body; (**d**) interaction with the TMP patches.

2.3. The ID Tracking Module

Being divided in two separated acquisition systems that work in parallel (i.e., SA for visual-auditory acquisition and TMP for physiological parameters acquisition), we needed a solution

to unify the data extracted from the subjects interacting with the robot in a unique meta-scene. This is mandatory in order to have a clear classification and association of the information with specific and permanent IDs.

The first issue to solve was a method that could guarantee the permanent recognition of IDs to be associated to the subjects of our interest. The Kinect has its own method for ID assignment, but it is unusable for our purpose, because every ID is assigned randomly and changes every time a subject disappears from and comes back in the field of view of the camera. This happens even if the detection is lost for a few seconds, for example, when a subject passes in front of another. There are a lot of options for subject recognition. There are published papers and patents about methods for recognizing people from their face [52–54] or their voice [55]. Nonetheless, these methods are time consuming, they often need a pre-defined data-set of faces and voices, and work rather well only when they are trained on the few persons who are most frequently exposed to these vision or audio systems. These techniques cannot be suitable solutions in an unpredictable and non-structured environment such as a social scenario, in which the subjects interacting with the robot are numerous, and potentially unknown until they appear in the vision system or they speak with the synthetic agent. Therefore, we explored several non-invasive available solutions that could be suitable for our scope and not expensive in terms of computational costs. Some of these techniques include NFC (Near Field Communication), RFID (Radio Frequency Identification) or Bluetooth devices, but we did not want to add any hardware device on the user side. For these reasons, we implemented the *QR Code Module* (QRCm) and integrated it in the SA framework (Figure 2). QRCm is one of the most recent modules of the SA, added to let the software recognise QR codes that can be printed as stickers and easily applied to the clothes of the users. The Microsoft HD camera, integrated in the robot body and replaceable with any HD cameras used for VoIP, is in charge for the related image acquisition. The generated QR codes are recognised as incremental natural numbers representing the permanent IDs associated with subjects. These IDs act like pointers to data stored in a long-term memory. Indeed, QRCm is connected with a dedicated external SQL (Structured Query Language) Database. The module can read, to check if a subject has already been detected in past interactions, or write new information in the database, such as names, physiological data and some individual anthropometric features of the interlocutors that make the recognition even more reliable. This choice corresponds to a small software addition exploiting the already available hardware on the robot, without the addition of any electrical device on the user side.

Nevertheless, being both the sensors for the SA and the TMP physically embodied in the robot, we encountered a secondary issue: almost all the times a person was requested to place their fingers on the android's shoulders in order to acquire physio data, that person approached the robot getting so close to it that the vision system was no longer able to detect any ID, not even by means of the QRCm, and the perceptual framework was not able to assign the gathered information to anyone. To solve this practical issue, we implemented the *ID Tracking Module* (IDTm).

The IDTm acts as a bridge between the SA and the TMP. Its mechanism is represented in the schema of Figure 7. Every time a subject approaches the patches covers most of the image, as well as the other people present in the scene. However, we have another consequence: the image becomes almost the same, especially the central part of it, regardless light conditions, contrast and brightness, and this reflects in a decrease of the variability of the pixels. This module continuously calculates the standard deviation of the pixels extracted by a cropped, re-sized and desaturated central part of the image captured by the Kinect camera. At the same time, SA provides 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, IDTm retrieves the ID of the last detected closest subject and, when the pixels standard deviation of the image comes under a threshold (empirically decided as 50), the module saves the ID of the last detected closest subject, assumes that he/she is the one approaching, and will assign potential physiological parameters detected by the TMP to that specific ID. The extracted physio data are stored in the database, assigned to the ID indicated by the

IDTm, along with all the other information about that person provided by the SA. Finally, as soon as the subject leaves the patches and comes back visible in the camera field of view, will be detected again by the SA and the QRCm.

This method leads to a novel concept for the database, which becomes a storage of many extended meta-scenes that include data from both SA and TMP, classified by time and IDs. This technique gives the possibility to estimate not only a temporary affective state, but also to do comparisons among different subjects in the same scenario, or between past and future acquisitions about the same person.

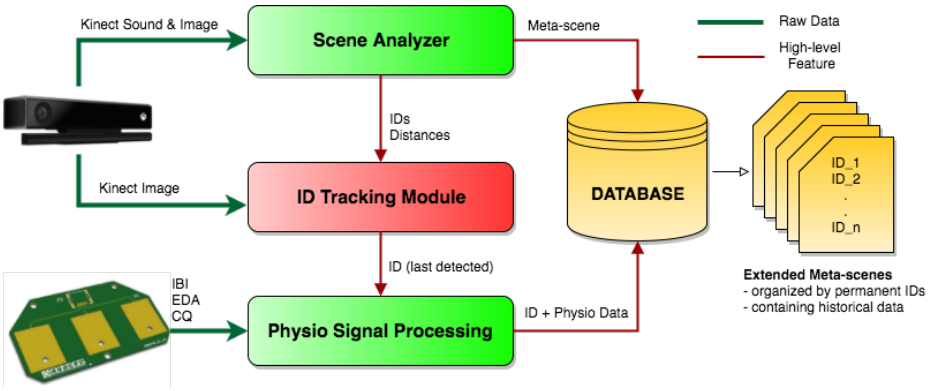

Figure 7. Flowchart of the subjects information flow managed by the "ID Tracking Module". This schema reports how the subject ID is detected and maintained, even in case of camera occlusion, thanks to the synergy between the two different acquisition systems modules: "Scene Analyzer" and "Physio Signal Processing".

3. Experiments

In this section, we present three experiments: the first one is an evaluation of the Touch-Me Pad, conducting a technical comparison with one of the best available device for physiological parameters acquisition; the second is about the presented IDTm module, therefore, is a demonstration of the capability of the system to keep track of the ID of subjects, even when they become undetectable for the vision system; the third and latter is the test of the overall architecture, included data integration, demonstrating the ability of our perception framework to gather data from both SA and TMP in a continuous session of acquisition.

3.1. Validation of the TMP

The performance of the developed TMP has been evaluated by comparing data acquired at the chest with data acquired at the fingers. In particular, the TMP system was compared with Biopac instrument (MP35 device from Biopac Systems Inc., Goleta, CA, USA), considered as a gold standard for the acquisition of physiological signal. Seven users were asked to attach Biopac Ag/AgCl electrodes in the thorax area and then to put their fingers in the TMP sensor in order to acquire ECG signals from both systems at the same time, during static position. Figure 8 shows the collected signals after being time aligned and then elaborated using a dedicated software tool (WFDB) available from Physionet (https://physionet.org/physiotools/matlab/). This tool could be used within Matlab software and allows to extract useful parameter from ECG signals [56,57]. In particular, we used the feature called *wqrs*, that attempts to locate QRS complexes in an ECG signal, in order to evaluate if it is possible to extract robust HR parameter also from signals gathered from user fingers.

Then, a Bland-Altman method was used to quantify the maximum error of the TMP on extracting the HR value from the IBI parameter. This method allows to statistically compare two signals, using a graphical representation.

In this graphical method, the differences between the two techniques are plotted against the averages of the two techniques. In the plots, horizontal lines represent respectively the

mean of the difference of the two systems and lower and upper *Limits of agreements*. If the 95% of the difference value remain within the *Limits of agreement*, the two systems of signals could be considered statistically equivalent. Considering the Bland-Altman plot reported in Figure 9, it is possible to underline that the TMP system is a valid instrument for physiological signals evaluation, with a maximum error in HR evaluation of ± 5 bpm.

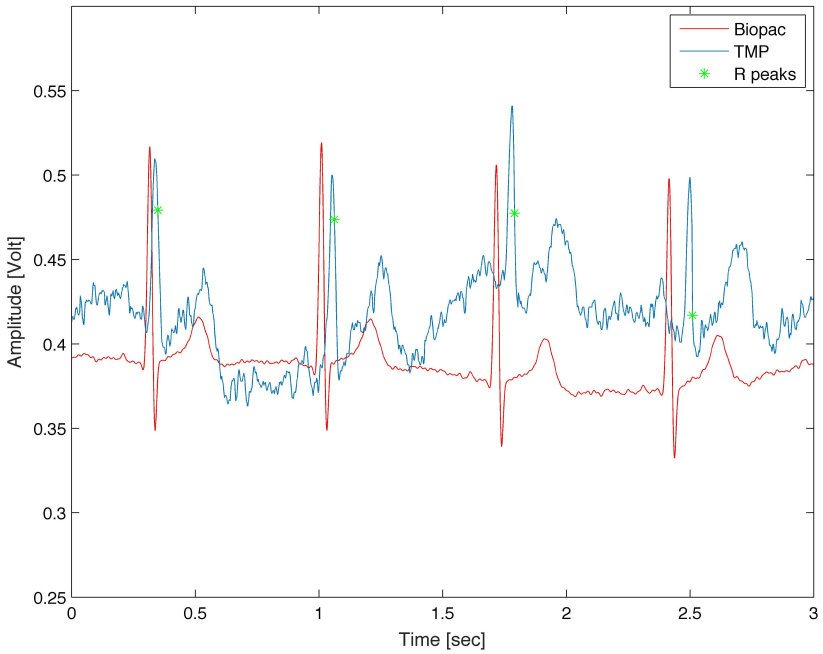

Figure 8. Synchronised ECG signals acquired from thorax region with Biopac instrument (red line) and from finger site with TMP system (blue line). TMP signal is processed with WFDB tool and R peaks detected are marked with green * symbols.

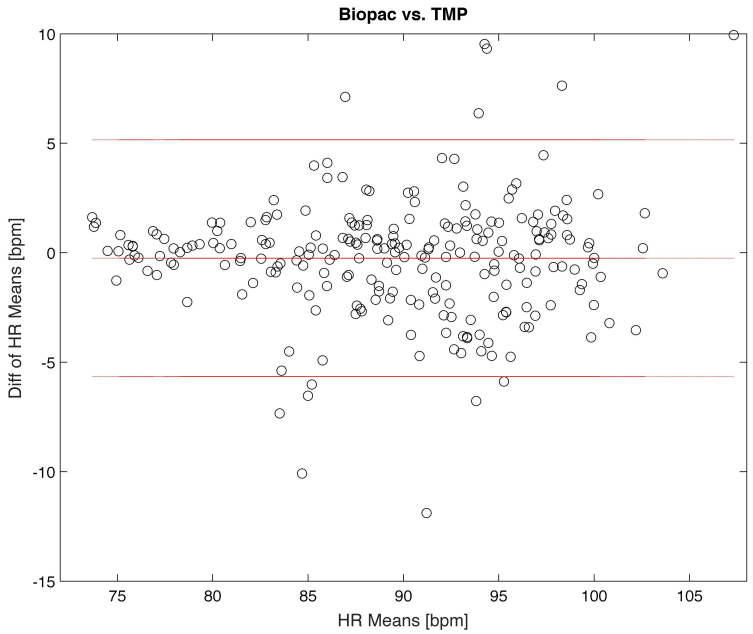

Figure 9. Bland-Altman plots of the Heart Rate value extracted from the TMP signal compared with the one calculated from Biopac signal.

3.2. Testing the IDTm Software

This experiment is focused on the capability of the IDTm to keep track of the users while monitoring their psycho-physiological state. With this purpose, two are the main parameters to be investigated: the *Pixels Standard Deviation* of the image captured by the camera and the *Subjects Position*, in terms of relative distance between the subjects and the vision sensor.

To stress the acquisition system, we selected the following scenario: two persons, a male and a female, distinguished by two different QR codes applied on their clothes, are at a similar initial distance from the robot. The male will approach the robot and place his fingers on the TMP contacts for the physio data extraction, which will last the minimum time necessary to let the TMP acquire reliable IBI and EDA signals (60 s). During the gathering of the physiological data he will occlude the vision sensor, however, the system must be aware of who is touching the patches. Moreover, when the subject moves away from the robot, the perceptual framework has to be able to recognise again and differentiate the two subjects with their previous IDs.

In Figure 10 we report the results of this experiment. The upper chart shows the trend of the pixels standard deviation of a cropped central portion of the image captured by the Kinect (green line). The minimum threshold at 50 has been highlighted with a straight dotted purple line. In the lower chart, it is possible to see the relative distances between each of the two subjects and the vision sensor. They are a 26 years old female having a QR code correspondent to ID 1 (red line in the line chart) and a 27 years old male, whose ID is 2 (blue line).

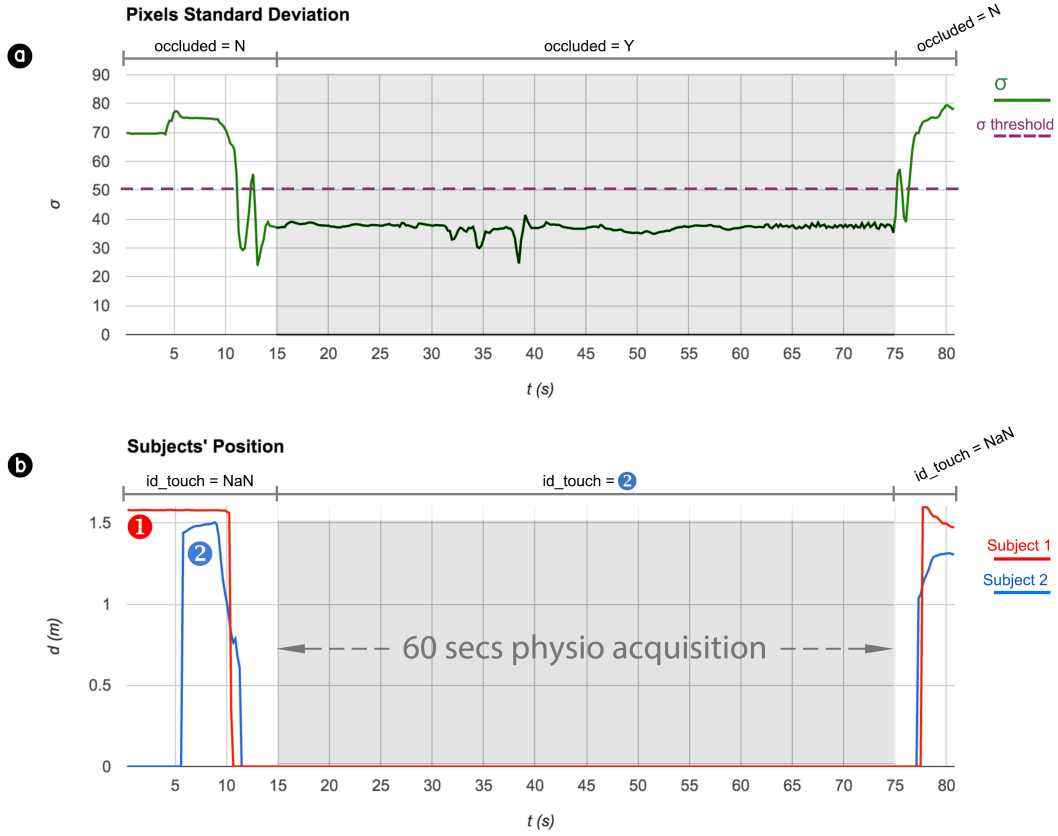

Figure 10. IDTm (ID Tracking Module) Validation Test; (**a**) Pixels Standard Deviation of the elaborated image; (**b**) position of the subjects in terms distance (m).

Some events are labelled in both charts. For example, at the beginning, and for the first 15 s, the robot has the possibility to see and analyse the social scene in which is involved, the fact that the camera is not occluded is given by the high standard deviation of the pixels given by a highly

variable image captured by the vision system. In the scene, both subjects are detected: subject 1, whose QR code is recognised from the beginning, stands at 1.6 m from the robot; subject 2 is recognised at 5.3 s and stands at 1.5 m from the robot. After a few seconds, subject 2 approaches the FACE robot and his detected distance decreases. At $t = 11$ s subject 2 is so close to the robot that covers the view of subject 1, who is lost by the vision system ($d = 0$). After only one second, even subject 2, who is approaching, exceeds the minimum distance needed by the SA to detect and recognise subjects, and gets lost by the vision system. At $t = 15$ s, subject 2 has reached the electronic patches of the TMP occluding the camera enough to let the pixel standard deviation goes under its minimum threshold. Here starts the 60 s acquisition of physiological parameters (from $t = 15$ s to $t = 75$ s). Finally, subject 2 removes his hands from the robot's shoulders and moves backwards.

During the experiment, we noticed two fundamental results: (1) although the two subjects got lost by the SA at a distance of one second only, the IDTm managed to successfully estimate who was touching the TMP patches, as the last closest subject's ID that has been detected. Indeed, during the physio data acquisition, the IDTm variable id_{touch} was set at 2, and this was correctly the ID to which the gathered signals had to be assigned; (2) only a few seconds after the physio data acquisition, thanks to the increase of the pixels' standard deviation over the threshold, the system realised that the camera was no longer occluded (occluded $= N$) and no one was touching the patches ($id_{touch} = NaN$), but what is more important is that, when the scene came back clear to the vision system, QRCm managed to recognise again both subjects assigning the two correct IDs.

3.3. Integrating the Data

The aim of this last experiment is to demonstrate the multimodal interaction of the overall perceptual framework. We run several experiments with two males and two females acquiring audio, visual and touch input, extracting high-level social information about gestures, facial expressions, proximity and psycho-physiological states of the interlocutors.

In Figure 11 an extraction of physio data (i.e., IBI, EDA and CQ), together with estimated facial expression divided per percentage ratios (i.e., anger ratio, happiness ratio, suprise ratio and sadness ratio), and distance from the vision sensor, is presented. The subject was a 32 years old male performing a sequence of fundamental expressions and alternatively approaching and touching the TMP sensors for acquiring physiological data, in a single session of approximately 4 min. All data have been extracted from the same extended meta-scene and related to the unique ID assigned to that subject.

The presented chart reveals a fundamental result. In many parts of the experiment, it can be noticed that the detection of the subject by the vision system is lost ($d = 0$). In the same manner, the acquisition of physiological parameters is discontinuous. This is particularly evident during physio data acquisition. It depends on the very close distance required for the physio extraction, and the consequence is a freezing or a loss of the information provided by the SA, i.e., facial expression in the first chart, and distance in the last one. Nonetheless, the perceptual interface manages to keep track of the ID of that subject and to acquire and store data, even when one of the two systems looses the acquisition flow. This is due to the overall framework, which is in charge of supervising and managing the acquisition in a similar way the human perception system does, thanks to the synergy among different senses.

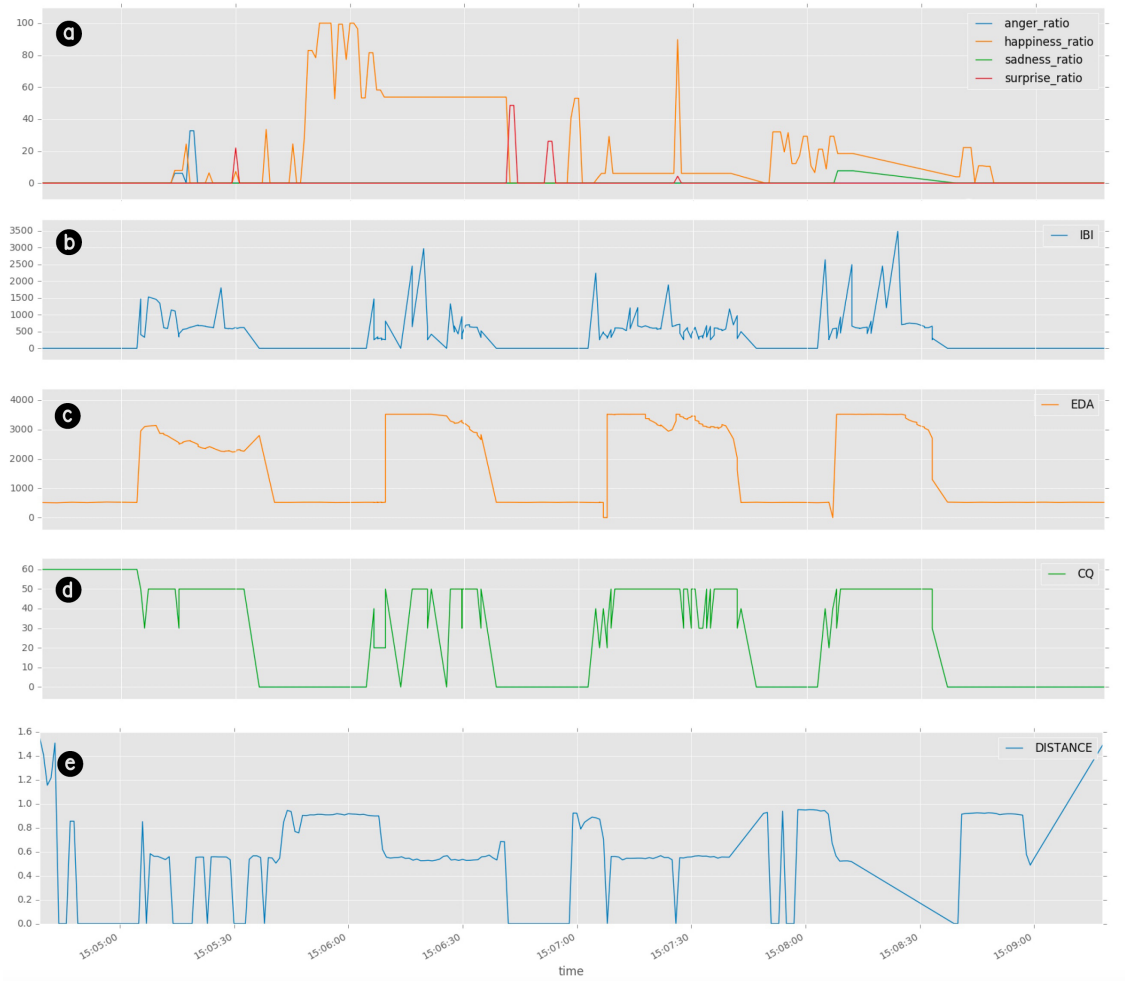

Figure 11. Data extracted from the database about: (**a**) facial expression estimation (expression ratio); (**b**) Inter-Beat Interval (IBI); (**c**) Electro Dermal Activity (EDA); (**d**) Contact Quality (CQ); (**e**) distance of the subject (in meters)

4. Discussion

The maintenance of unique individual IDs for different subjects, demonstrated in the first experiment, allows the usage of the presented perceptual framework for making comparisons among different subjects present in the same scene, but also to compare data of different sessions performed with the same subject. This property, coupled with the ability to extract social information from natural interaction scenarios, makes this system a particularly suitable tool for social robotics. Especially in that kind of applications in which a long-term study is needed, such as in educational contexts, where having the possibility to perform personalised learning is mandatory [58]. With this system it is possible to assign, store and retrieve permanent information about interlocutors in a fully automatic manner without invalidating the interaction. On the contrary, everything in the software and the hardware design has been studied in order to elicit social empathic interactions. The choice of replacing wearable sensors with embodied electronic patches that the user is sporadically asked to touch, as well as the simplification of time and computational costs in the acquisition, are just two examples of how engaging and human-centered this perceptual framework has been conceived. The extraction of multimodal inputs and their effective integration in a unique data structure, reported in the second experiment, increases the potential of the perception giving the chance to cognitive systems to infer more reliable social meanings about human behaviors and to disambiguate. For instance, a gesture captured by the presented perceptual interface can be confirmed by a physiological acquisition or

contradicted by inconsistent physiological signals. On the other hand, a peak in the electrodermal activity, revealed by the TMP, may have a qualitative value integrating this data with the estimation of the facial expression provided by the SA.

5. Future Work

The presented perception system needs to be tested applying several variations, e.g., stressing the occlusion occurrence, retesting with multiple subjects and multiple repetition of the presented experiment. Although the system has been already validated, these further experiments will lead to more robust results. The perceptual framework is already used in the context of the EASEL European Project (http://easel.upf.edu/), from which has been partially founded, and several experiments are in progress. In EASEL, this acquisition system is integrated in different humanoid robots that assume the role of synthetic tutors in educational scenarios (e.g., museums, school classrooms). In such environments, the real time estimation of pupils emotional state is mandatory for the modulation of the educational content delivered by the robots. In any case, being completely modular and portable, it would be very interesting for the future to consider the exploitation of our perception framework also out of social robotics context. For example, in case of humans and robots collaboration in production lines and other work related tasks [59–61]. In these applications, a continuous monitoring of the human psycho-physiological state by the robots can drastically decrease the risk of accidents and maximise productivity.

6. Conclusions

In this work, an unobtrusive perception framework for social robots has been presented and tested. The system has been designed exploiting the synergy between two main acquisition systems: the audiovisual acquisition system called Scene Analyzer, and the electronic patches for physiological signal extraction called Touch-Me Pad. In order to maximise the naturalness of the social interaction, and according to the importance of embodiment, all the sensory devices have been physically integrated in the FACE humanoid robot, an hyper-realistic android with which we performed two experiments. With these experiments, we tested the maintenance of the tracking of the subjects during the human-robot interaction regardless which part of the perceptual interface was in charge for the acquisition, and the effective integration of all the gathered information in a unique data-set called meta-scene. A meta-scene can be considered as the output of our perception framework and the input for potential cognitive systems that thanks to the presented framework will have the chance to analyse a complete collection of data about the social environment in which the social robot is involved. This will lead to an important improvement in social robotics, making the human behavior understanding more reliable even in crowded and noisy environment and in applications where long-term interactions have to be studied.

Acknowledgments: This work is partially supported by grants from the European Research Council under the European Union's 7th Framework Programme FP7/2007-2013/ERC grant agreement n. 611971 (EASEL) and n. 341196 (CDAC) to Paul F. M. J. Verschure.

Author Contributions: "Scene Analyzer" and FACE Robot have been developed by FACE Team (Daniele Mazzei, Lorenzo Cominelli and Roberto Garofalo); "Touch-Me Pad" has been developed by Alessandro Tognetti and Nicola Carbonaro; Lorenzo Cominelli, Nicola Carbonaro, Daniele Mazzei and Alessandro Tognetti conceived the experiments; Lorenzo Cominelli, Nicola Carbonaro and Roberto Garofalo performed the experiments and analyzed the data; Lorenzo Cominelli wrote the paper; Danilo De Rossi supervised all the conducted experiments.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Damasio, A. *Descartes' Error: Emotion, Reason, and the Human Brain*; Grosset/Putnam: New York, NY, USA, 1994.
2. Cowie, R.; Douglas-Cowie, E.; Tsapatsoulis, N.; Votsis, G.; Kollias, S.; Fellenz, W.; Taylor, J.G. Emotion recognition in human-computer interaction. *IEEE Signal Process. Mag.* **2001**, *18*, 32–80.
3. Brave, S.; Nass, C. Emotion in human-computer interaction. In *Human-Computer Interaction*; CRC Press: Boca Raton, FL, USA; London, UK; New York, NY, USA, 2003; p. 53.
4. Schachter, S.; Singer, J. Cognitive, social, and physiological determinants of emotional state. *Psychol. Rev.* **1962**, *69*, 379–399.
5. Picard, R.W. Affective computing: Challenges. *Int. J. Hum.-Comput. Stud.* **2003**, *59*, 55–64.
6. Jaimes, A.; Sebe, N. Multimodal human-computer interaction: A survey. *Comput. Vis. Image Underst.* **2007**, *108*, 116–134.
7. Yang, M.H.; Kriegman, D.J.; Ahuja, N. Detecting faces in images: A survey. *IEEE Trans. Pattern Anal. Mach. Intell.* **2002**, *24*, 34–58.
8. Hjelmås, E.; Low, B.K. Face detection: A survey. *Comput. Vis. Image Underst.* **2001**, *83*, 236–274.
9. Zhao, W.; Chellappa, R.; Phillips, P.J.; Rosenfeld, A. Face recognition: A literature survey. *ACM Comput. Surv. (CSUR)* **2003**, *35*, 399–458.
10. Fasel, B.; Luettin, J. Automatic facial expression analysis: A survey. *Pattern Recognit.* **2003**, *36*, 259–275.
11. Pantic, M.; Rothkrantz, L.J.M. Automatic analysis of facial expressions: The state of the art. *IEEE Trans. Pattern Anal. Mach. Intell.* **2000**, *22*, 1424–1445.
12. Pierre-Yves, O. The production and recognition of emotions in speech: Features and algorithms. *Int. J. Hum.-Comput. Stud.* **2003**, *59*, 157–183.
13. Murray, I.R.; Arnott, J.L. Toward the simulation of emotion in synthetic speech: A review of the literature on human vocal emotion. *J. Acoust. Soc. Am.* **1993**, *93*, 1097–1108.
14. Marcel, S. *Gestures for Multi-Modal Interfaces: A Review*; Technical Report; IDIAP: Martigny, Switzerland, 2002.
15. Pavlovic, V.I.; Sharma, R.; Huang, T.S. Visual interpretation of hand gestures for human-computer interaction: A review. *IEEE Trans. Pattern Anal. Mach. Intell.* **1997**, *19*, 677–695.
16. Hu, W.; Tan, T.; Wang, L.; Maybank, S. A survey on visual surveillance of object motion and behaviors. *IEEE Trans. Syst. Man Cybern. Part C (Appl. Rev.)* **2004**, *34*, 334–352.
17. Wang, J.J.; Singh, S. Video analysis of human dynamics—A survey. *Real-Time Imaging* **2003**, *9*, 321–346.
18. Gavrila, D.M. The visual analysis of human movement: A survey. *Comput. Vis. Image Underst.* **1999**, *73*, 82–98.
19. Potamianos, G.; Neti, C.; Luettin, J.; Matthews, I. Audio-visual automatic speech recognition: An overview. In *Issues Visual and Audio-Visual Speech Processing*; MIT Press: Cambridge, MA, USA, 2004; Volume 22, p. 23.
20. Duchowski, A.T. A breadth-first survey of eye-tracking applications. *Behav. Res. Methods Instrum. Comput.* **2002**, *34*, 455–470.
21. Ji, Q.; Wechsler, H.; Duchowski, A.; Flickner, M. Special Issue: Eye Detection and Tracking. *Comput. Vis. Image Underst.* **2005**, *98*, 210.
22. Porta, M. Vision-based user interfaces: Methods and applications. *Int. J. Hum.-Comput. Stud.* **2002**, *57*, 27–73.
23. Duric, Z.; Gray, W.D.; Heishman, R.; Li, F.; Rosenfeld, A.; Schoelles, M.J.; Schunn, C.; Wechsler, H. Integrating perceptual and cognitive modeling for adaptive and intelligent human-computer interaction. *Proc. IEEE* **2002**, *90*, 1272–1289.
24. Turk, M.; Kölsch, M. Perceptual interfaces. In *Emerging Topics in Computer Vision*; Prentice Hall: Upper Saddle River, NJ, USA, 2004.
25. Tognetti, S.; Garbarino, M.; Bonarini, A.; Matteucci, M. Modeling enjoyment preference from physiological responses in a car racing game. In Proceedings of the 2010 IEEE Symposium on Computational Intelligence and Games (CIG), Copenhagen, Denmark, 18–21 August 2010; pp. 321–328.
26. Gilleade, K.; Dix, A.; Allanson, J. Affective videogames and modes of affective gaming: Assist me, challenge me, emote me. In Proceedings of the DiGRA 2005: Changing Views—Worlds in Play, Vancouver, BC, Canada, 16–20 June 2005.
27. Montalbana, I.; Leona, A.G.E. Emotion-aware intelligent environments: A user perspective. In *Intelligent Environments 2009*; IOS Press: Amsterdam, Netherlands, 2009; p. 421.

28. Liu, C.; Conn, K.; Sarkar, N.; Stone, W. Physiology-based affect recognition for computer-assisted intervention of children with Autism Spectrum Disorder. *Int. J. Hum.-Comput. Stud.* **2008**, *66*, 662–677.
29. Meltzoff, A.N. The 'like me' framework for recognizing and becoming an intentional agent. *Acta Psychol.* **2007**, *124*, 26–43.
30. Fong, T.; Nourbakhsh, I.; Dautenhahn, K. A survey of socially interactive robots. *Robot. Auton. Syst.* **2003**, *42*, 143–166.
31. Breazeal, C. Emotion and sociable humanoid robots. *Int. J. Hum.-Comput. Stud.* **2003**, *59*, 119–155.
32. Broekens, J.; Heerink, M.; Rosendal, H. Assistive social robots in elderly care: A review. *Gerontechnology* **2009**, *8*, 94–103.
33. Sharkey, A.; Sharkey, N. Granny and the robots: Ethical issues in robot care for the elderly. *Ethics Inf. Technol.* **2012**, *14*, 27–40.
34. Scassellati, B.; Admoni, H.; Matarić, M. Robots for use in autism research. *Annu. Rev. Biomed. Eng.* **2012**, *14*, 275–294.
35. Cabibihan, J.J.; Javed, H.; Ang, M.; Aljunied, S.M. Why robots? A survey on the roles and benefits of social robots in the therapy of children with autism. *Int. J. Soc. Robot.* **2013**, *5*, 593–618.
36. Robins, B.; Dautenhahn, K.; Te Boekhorst, R.; Billard, A. Robotic assistants in therapy and education of children with autism: Can a small humanoid robot help encourage social interaction skills? *Univers. Access Inf. Soc.* **2005**, *4*, 105–120.
37. Saerbeck, M.; Schut, T.; Bartneck, C.; Janse, M.D. Expressive robots in education: Varying the degree of social supportive behavior of a robotic tutor. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, Atlanta, GA, USA, 10–15 April 2010; ACM: New York, NY, USA, 2010; pp. 1613–1622.
38. Causo, A.; Vo, G.T.; Chen, I.M.; Yeo, S.H. Design of Robots Used as Education Companion and Tutor. In *Robotics and Mechatronics*; Springer: Cham, Switzerland, 2016; pp. 75–84.
39. 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.* **2014**, *44*, 157–168.
40. Metta, G.; Fitzpatrick, P.; Natale, L. YARP: Yet another robot platform. *Int. J. Adv. Robot. Syst.* **2006**, *3*, 43–48.
41. Zaraki, A.; Pieroni, M.; De Rossi, D.; Mazzei, D.; Garofalo, R.; Cominelli, L.; Dehkordi, M.B. Design and Evaluation of a Unique Social Perception System for Human-Robot Interaction. *IEEE Trans. Cogn. Dev. Syst.* **2016**, *PP*, doi:10.1109/TCDS.2016.2598423.
42. Lisetti, C.L.; Nasoz, F. Using Noninvasive Wearable Computers to Recognize Human Emotions from Physiological Signals. *EURASIP J. Adv. Signal Process.* **2004**, *2004*, 929414.
43. Venables, P.H.; Christie, M.J. Electrodermal activity. In *Techniques in Psychophysiology*; John Wiley: Hoboken, NJ, USA, 1980; Volume 54.
44. Poh, M.Z.; Swenson, N.C.; Picard, R.W. A wearable sensor for unobtrusive, long-term assessment of electrodermal activity. *IEEE Trans. Biomed. Eng.* **2010**, *57*, 1243–1252.
45. Kappeler-Setz, C.; Gravenhorst, F.; Schumm, J.; Arnrich, B.; Tröster, G. Towards long term monitoring of electrodermal activity in daily life. *Pers. Ubiquitous Comput.* **2013**, *17*, 261–271.
46. 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 Proceedings of the GLOBAL HEALTH 2012: The First International Conference on Global Health Challenges, Venice, Italy, 21–26 October 2012.
47. 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 Proceedings of the 2011 IEEE International Symposium on a World of Wireless, Mobile and Multimedia Networks (WoWMoM), Lucca, Italy, 20–24 June 2011; pp. 1–6.
48. Wainer, J.; Feil-Seifer, D.J.; Shell, D.A.; Mataric, M.J. The role of physical embodiment in human-robot interaction. In Proceedings of the 15th IEEE International Symposium on Robot and Human Interactive Communication (ROMAN 2006), Hatfield, UK, 6–8 September 2006; pp. 117–122.
49. Jung, Y.; Lee, K.M. Effects of physical embodiment on social presence of social robots. In Proceedings of the PRESENCE, Valencia, Spain, 13–15 October 2004; pp. 80–87.
50. Mazzei, D.; Billeci, L.; Armato, A.; Lazzeri, N.; Cisternino, A.; Pioggia, G.; Igliozzi, R.; Muratori, F.; Ahluwalia, A.; De Rossi, D. The face of autism. In Proceedings of the 2010 IEEE on RO-MAN, Viareggio, Italy, 13–15 September 2010; pp. 791–796.

51. Mazzei, D.; Lazzeri, N.; Hanson, D.; De Rossi, D. HEFES: An hybrid engine for facial expressions synthesis to control human-like androids and avatars. In Proceedings of the 2012 4th IEEE RAS & EMBS International Conference on Biomedical Robotics and Biomechatronics (BioRob), Rome, Italy, 24–27 June 2012; pp. 195–200.
52. Haritaoglu, I.; Harwood, D.; Davis, L.S. W4S: A real-time system for detecting and tracking people in 2 1/2D. In *Computer Vision—Proceedings of the ECCV'98: 5th European Conference on Computer Vision, Freiburg, Germany, 2–6 June 1998*; Burkhardt, H., Neumann, B., Eds.; Springer: Berlin/Heidelberg, Germany, 1998; Volume I, pp. 877–892.
53. Brunelli, R.; Falavigna, D.; Poggio, T.; Stringa, L. Recognition System, Particularly for Recognising People. U.S. Patent 5,412,738, 1995.
54. Turk, M.; Pentland, A. Face Recognition System. U.S. Patent 5,164,992, 17 November 1992.
55. Doddington, G.R. Speaker recognition? Identifying people by their voices. *Proc. IEEE* **1985**, *73*, 1651–1664.
56. Silva, I.; Moody, G.B. An open-source toolbox for analysing and processing physionet databases in matlab and octave. *J. Open Res. Softw.* **2014**, *2*, e27, doi:10.5334/jors.bi.
57. Goldberger, A.L.; Amaral, L.A.; Glass, L.; Hausdorff, J.M.; Ivanov, P.C.; Mark, R.G.; Mietus, J.E.; Moody, G.B.; Peng, C.K.; Stanley, H.E. Physiobank, physiotoolkit, and physionet. *Circulation* **2000**, *101*, e215–e220.
58. Sampson, D.; Karagiannidis, C. Personalised learning: Educational, technological and standardisation perspective. *Interact. Educ. Multimed.* **2010**, 24–39.
59. Makris, S.; Karagiannis, P.; Koukas, S.; Matthaiakis, A.S. Augmented reality system for operator support in human-robot collaborative assembly. *CIRP Ann.-Manuf. Technol.* **2016**, *65*, 61–64.
60. Claes, D.; Tuyls, K. Human Robot-Team Interaction. In Proceedings of the Artificial Life and Intelligent Agents Symposium, Bangor, UK, 5–6 November 2014; Springer: Cham, Switzerland, 2014; pp. 61–72.
61. Tsarouchi, P.; Makris, S.; Chryssolouris, G. Human-robot interaction review and challenges on task planning and programming. *Int. J. Comput. Integr. Manuf.* **2016**, *29*, 916–931.