Exploring Human attitude during Human-Robot Interaction

Exploring Human attitude during Human-Robot Interaction*

A.Sorrentino, L. Fiorini, *Member, IEEE,* I.Fabbricotti, D.Sancarlo, F.Ciccone, F.Cavallo*, Member IEEE*

Abstract— **The aim of this work is to provide an automatic analysis to assess the user attitude when interacts with a companion robot. In detail, our work focuses on defining which combination of social cues the robot should recognize so that to stimulate the ongoing conversation and how. The analysis is performed on video recordings of 9 elderly users. From each video, low-level descriptors of the behavior of the user are extracted by using open-source automatic tools to extract information on the voice, the body posture, and the face landmarks. The assessment of 3 types of attitude (neutral, positive and negative) is performed through 3 machine learning classification algorithms: k-nearest neighbors, random decision forest and support vector regression. Since intra- and inter-subject variability could affect the results of the assessment, this work shows the robustness of the classification models in both scenarios. Further analysis is performed on the type of representation used to describe the attitude. A raw and an auto-encoded representation is applied to the descriptors. The results of the attitude assessment show high values of accuracy (>0.85) both for unimodal and multimodal data. The outcome of this work can be integrated into a robotic platform to automatically assess the quality of interaction and to modify its behavior accordingly.**

I. Introduction

Companion robots will permeate our daily life in a near future thus they were required to show a high level of social interaction. During a human-robot interaction (HRI), companion robots were perceived as social actors and, consequently, they evoke mental models typical of a human-human interaction [1]. Social relationships are complex and include several cues such as the language, the tone, the emotion, the body posture, and the facial expression. Additionally, the attitude of a person towards social interaction is expressed by a set of social signals which conveys information about mental state, feelings and other personal traits (i.e. eye gazing, postures, voice quality) [2].

To overcome the HRI gaps, future social robots require to integrates cognitive models able to create a solid vision of how our interaction patterns work as underlined in this recent review paper [3]. They should be able to create stimulating and engaging interactions in which a user actively participates for an extended period of time. In order to achieve it, it is essential to identify which is the behavior of the user and how it changes during the interaction to shape the behavior of the robot accordingly. A first attempt in this direction has been described in [4], in which a NAO robot is endowed with the capability of simulating empathic behavior based on the recognized user emotion. The emotional behavior of the user has been assessed by mean of facial expression and speech prosody. Affective and interactive signals based on facial expressions are also provided as input (with tactile stimuli) to the socially adaptable framework of iCub robot in [5].

The scientific rationale behind this paper is to enrich the perceptual module, enlarging the set of features that can be used to determine the levels of engagement and the attitude during the interaction. The work described in [6] shows the importance of focusing on the robot perceptual step for assessing the engagement level in children with autism. The innovation of our work focuses the attention on the perceptual module of a companion robot with interacts with the elderly. In our experimental scenario, the elderly user engages in conversation with the robot, expressing multiple social cues that can be used to assess the engagement state. In this work, we identified the social cues that can be extracted by common sensors mounted over robotics platforms, listed in Table I, and we analyze different combinations of cues so that to explore which could be used by the behavioral model of the robot.

A. Study Design

The main goal of this analysis is to detect whether traditional machine learning algorithms can be used to assess the attitude of each user independently (Intra - classification task). Secondly, this paper aims to test the robustness of the proposed classification models due to the inter-subjects variability which could affect the results of the recognition task. Additionally, to evaluate which set of cues are more accurate, comparison with different datasets (Video and Audio/Video) was also performed. In order to achieve the proposed goal, the general emotions felt by the elderly were divided into three attitude spheres: neutral, positive and negative. In this work, a neutral attitude includes moments in which the elderly express a blended set of emotions ranging

* This work was supported by ACCRA Project, founded by the European Commission- Horizon 2020 Funding Programme (H2020-SCI-PM14-2016) and National Institute of Information and Communications Technology (NICT) of Japan under grant agreement No. 738251.

A. Sorrentino, L. Fiorini and F.Cavallo are with Scuola Superiore Sant'Anna, Pisa, Italy and with Department of Excellence in Robotics & AI, Piazza Martiri della Libertà, 33 - 56127 Pisa, Italy (corresponding author Alessandra Sorrentino: phone: +39 050 883478; e-mail: alessandra.sorrentino@ santannapisa.it).

I. Fabbricotti is with the Erasmus School of Health Policy and Management, Erasmus University, Rotterdam, The Netherlands.

D. Sancarlo and F. Ciccone are with the complex Unit of Geriatrics, Department of Medical Sciences, Fondazione "Casa Sollievo della Sofferenza" – IRCCS, San Giovanni Rotondo, Foggia, Italy.

F. Cavallo is with the Department of Industrial Engineering, University of Florence, Florence, Italy.

Manuscript 108 submitted to 2020 29th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN). Received April 9, 2020.

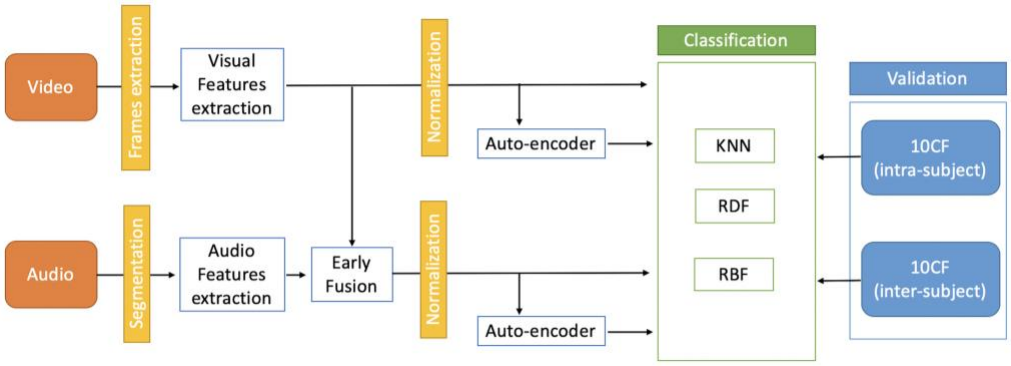

Figure 1. Data Analysis

TABLE I. SOCIAL CUES ANALYSED

Behavioral Aspect	Social Cue	Features	Sensor
Engagement state	Body Posture	Body orientation	Camera
	Head orientation	Roll, Pitch, Yaw angles	Camera
Emotional state	Expression	Facial action units	Camera
	Voice quality	Low Level Descriptors (i.e.Tempo,Energy,Pitch)	Microphone

from awaiting, expressionless and surprises. The positive attitude is characterized by emotional states that express joy, amusement, happiness, and affection. The negative attitude is characterized by elderly behavior expressing disappointment, irritation, annoyance, and impatience.

The performance of machine learning methods is heavily dependent on the choice of data representation (or features) on which they are applied [7]. For that reason, we compare the chosen classifiers on two different types of representations: raw and auto-encoded [8]. The auto-encoded representation is obtained as the output of the encoder layer. It is characterized by a reduced dimension with respect to the raw representation, composed only by the most relevant features. The choice of using an auto-encoder is to reduce the problem of noise and partially observed data [6].

II. EXPERIMENTAL SETTING

A. Participants.

A total of 9 users (8 female, 1 male user, avg age=83.55 years old, std age=4 years old) were enrolled for this study. None of them had hearing or speaking impairments. All the participants signed the consent form before entering the test. All the tests took place at IRCSS Casa Sollievo della Sofferenza (San Giovanni Rotondo, FG, Italy). The study was conducted in accordance with the Declaration of Helsinki, and the protocol was approved by the Ethics Committee of "Fondazione Casa Sollievo della Sofferenza" in San Giovanni Rotondo, Italy.

B. Buddy Robot

Buddy is a fully mobile robot moving with two motorized wheels, an articulated head and a plethora of sensors (i.e. 6 obstacle sensors, odometer, accelerometer, 7 ground sensors, RGB camera, 3D camera, thermal matrix, 3 caress sensors, 1 array of 4 microphones). Buddy is a robot designed for human-robot interaction. Its sensors can enhance its obstacles and cliff avoidance capability. This gives Buddy the capacity to safely look for a user. About its speaking capabilities, Buddy is able to understand specific questions in three languages: English, Dutch and Italian. This design choice excludes the possibility to interact freely with Buddy since there is no Dialog Manager included.

C. Experimental protocol

After initial training on the use of Buddy robot, the user was asked to individually interact with Buddy robot by asking it some general questions (i.e. "Hi Buddy, how are you?", "which is your favorite color?") and Buddy was supposed to answer like in a real conversation. A clinician participated in the interaction session as external support and he/she was ready to intervene in case of necessity.

The attitude assessment of the users during the experimental session has been off-line performed by the human expert of Erasmus University (NE). Namely, she annotated the intervals of time of the videos belonging to the three attitudes of the user: neutral, positive and negative attitude. The information reported at this stage is used as ground truth for the automatic attitude assessment.

Additionally, to assess the reliability of the system, the number of correct and incorrect answers, the repeated questions and the questions not understood were also off-line annotated by the expert.

III. DATA ANALYSIS

The interaction with Buddy robot has been evaluated by automatically analyzing the recorded video. Namely, from each video, the features listed in Table I are extracted from each frame. The features extracted from image frames have been collected into a unimodal dataset. The multimodal dataset has been obtained by augmenting each instance of the previous dataset with the audio data (early fusion). The attitude assessment has been performed on both datasets, by using the raw and the auto-encoded representation of each instance. The full data analysis process is depicted in Fig. 1.

A. Feature extraction

The recorded video was then analyzed to extract the visual and audio features. At the end of this process, two datasets were obtained: the first one contains the data

Manuscript 108 submitted to 2020 29th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN). Received April 9, 2020.

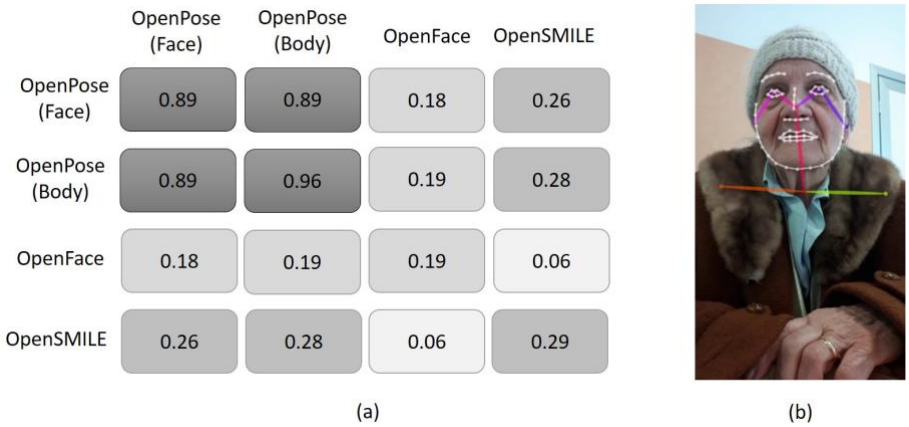

Figure 2. (a) The fraction of features detected by the extraction tools across the different modalities. (b) An example of keypoints extracted

extracted from the video (unimodal) and the second one contains the data extracted from the video and the audio (multimodal).

1) Visual Features Extraction

The visual features of interest have been extracted by each image frame (sampled at 30Hz) using two open-source data processing tools: OpenPose[9] and OpenFace [10].

OpenPose is a real-time multi-person 2D pose estimation framework. On a single image, this bottom-up approach jointly estimates the human body, foot, hand, and facial keypoints (in total 135 keypoints). As shown in [9], this approach exceeds the previous state-of-the-art results both in performance and efficiency. For this reason, a pre-trained version of the system has been adopted in this work. The framework has been tested over videos of dimension 1080 x 1920 pixels to extract the 2D position of 25 body joints and 70 facial keypoints. The list of estimated body keypoints has been filtered, removing the keypoints of the lower part of the body, because occluded during the experimental session, and of the hands, which are barely detected due to their closeness to the camera. As result, 30 poses of 8 body joints are extracted every second. The joints correspond to the nose, neck, shoulders, eyes, and ears.

In addition to the 70 facial keypoints extracted with the OpenPose tool, OpenFace toolkit is used to analyze the facial behavior of the user. In detail, we used the OpenFace toolkit to estimate head pose and facial action unit (AUs), which are commonly used to assess emotion in affective computing. As described in [10], the OpenFace model uses a Convolutional Experts Constrained Local Model (CE-CLM) which is composed of Point Distribution Model (PDM), which detect landmark shape variations, and patch experts, which model local appearance variations of each landmark. The estimated head pose is expressed in terms of the location of the head with respect to the camera in millimeters (Tx, Ty, Tz) and of rotations in radians around x,y,z axes (Rx, Ry, Rz). The 18 recognized facial action units are expressed in terms of their presence (0-1) and their intensity (on a 6 level Likert scale).

2) Audio Features Extraction

The acoustic low-level descriptors (LLDs) are extracted from the speech waveform on the frame level by using the open-source data processing tool OpenSMILE [11]. Specifically, we used this tool to extract 55 LLDs: RMS energy, Spectral absolute difference, Spectral flux, Spectral entropy, Spectral variance, Spectral skewness, Spectral kurtosis, Spectral slope, Spectral harmonicity, F0 (ACF based), F0 envelope (ACF based), unclipped voicing probability, Jitter(ACF based), Jitter DDP (ACF based), shimmer(ACF based), Logarithmic HNR, F0 (SHS based), F0 envelope (SHS based), unclipped voicing probability, Jitter (SHS based), Jitter DDP (SHS based), shimmer (SHS based), MFCC 0-14, Log Mel frequency band 0-7, LSP frequency 0-7, loudness and zcr. These features were computed over sliding windows of length 43 ms with a 33.33 ms shift and then aligned with the visual features using timestamps stored during the data recording.

B. Classification

We performed intra- and inter-subject validation of unimodal and multimodal data by using both raw and auto-encoded data. In the intrasubject case, the classification is performed on the features of each elder individually. To minimize the bias, the 10-Cross Fold validation technique was applied to each elderly dataset. The 10-Cross Fold validation technique was used also in the inter-subject classification, where the features of all the users were merged. By using the raw representation, unimodal data were represented by a vector of real-valued numbers $x = [x_{openPose}; x_{openFace}] \in \Re^{Dx1}$, where D=277 is the total number of visual features. No unimodal data of audio features were classified since the quantity of available data were extremely few. The raw representation of multimodal data was obtained augmenting the vector x with the audio features as $x= [x_{openPose}; x_{openFace}; x_{openSMILE}] \in \Re^{Dx1}$, resulting D=332. The z-normalization is applied on both on unimodal and multimodal vectors. According to the ground truth table provided by the expert, each instance was manually labeled.

The size of the auto-encoded space varied in order to fit the size of the input features. Since the main goal of the autoencoder is to compress the given input (encoder) and to reconstruct it from the compressed version (decoder), we first focused on the dimension of the compressed size. After different trials, we found out that the optimal encoded sizes were 128 and 200. It means that the decoder was able to acceptably reconstruct the input from the compressed representation provided by the encoder of dimension 128 and

TABLE III. CLASSIFIER PERFORMANCES (U= UNIMODAL DATASET, M=MULTIMODAL DATASET) FOR THE INTRA-SUBJECT (INTRA) AND INTER-SUBJECT (INTER) ANALYSIS

Algorithm	Representation	Modality	Accuracy		F-measure		Precision		Recall		Time	
			Intra	*Inter*	*Intra*	*Inter*	*Intra*	*Inter*	*Intra*	*Inter*	*Intra*	*Inter*
KNN	raw	U	0.971	0.970	0.941	0.968	0.947	0.969	0.947	0.967	3.884	235.409
		M	0.978	0.976	0.980	0.974	0.961	0.975	0.957	0.973	3.463	202.617
	128b_128e	U	0.954	0.960	0.941	0.957	0.945	0.956	0.940	0.956	1.012	14.602
		M	0.960	0.963	0.950	0.961	0.952	0.962	0.949	0.960	0.852	18.940
	128b_200e	U	0.941	0.959	0.919	0.956	0.925	0.957	0.922	0.956	1.031	17.153
		M	0.955	0.967	0.941	0.967	0.949	0.968	0.937	0.966	1.540	26.957
	200b_128b	U	0.955	0.962	0.930	0.959	0.935	0.960	0.929	0.959	1.704	25.009
		M	0.954	0.968	0.934	0.965	0.933	0.966	0.934	0.964	1.473	29.133
RDF	raw	U	0.988	0.986	0.980	0.985	0.986	0.986	0.974	0.985	5.463	99.074
		M	0.988	0.988	0.980	0.987	0.988	0.988	0.974	0.987	5.884	98.860
	128b_128e	U	0.949	0.956	0.929	0.953	0.948	0.955	0.915	0.952	3.515	59.167
		M	0.961	0.959	0.940	0.956	0.958	0.957	0.927	0.954	3.497	54.868
	128b_200e	U	0.949	0.955	0.922	0.952	0.950	0.954	0.905	0.951	3.734	59.978
		M	0.963	0.959	0.929	0.959	0.961	0.961	0.911	0.958	4.583	70.493
	200b_128b	U	0.952	0.957	0.920	0.954	0.953	0.957	0.900	0.953	4.505	63.005
		M	0.961	0.961	0.928	0.959	0.955	0.960	0.909	0.958	4.393	69.397
RBF	raw	U	0.915	0.891	0.866	0.883	0.902	0.891	0.805	0.878	11.516	1062.506
		M	0.901	0.891	0.871	0.890	0.912	0.899	0.779	0.885	11.850	1330.565
	128b_128e	U	0.851	0.860	0.787	0.850	0.850	0.857	0.767	0.846	7.704	519.113
		M	0.832	0.864	0.803	0.853	0.847	0.862	0.758	0.851	7.648	516.590
	128b_200e	U	0.855	0.862	0.796	0.852	0.849	0.859	0.735	0.848	7.507	511.187
		M	0.869	0.882	0.812	0.873	0.876	0.880	0.793	0.870	10.691	816.110
	200b_128b	U	0.838	0.869	0.755	0.859	0.846	0.867	0.708	0.855	10.934	742.650
		M	0.854	0.870	0.786	0.860	0.857	0.869	0.712	0.856	11.075	791.216

200. The loss of the autoencoder was minimized using the Adadelta gradient descent algorithm with learning rate equals to 1 and 200 epochs. With the aim of understanding which configuration held better results, this work compares the representations obtained with: an encoder of size 128 trained with the same batch size (128b_128e), an encoder of size 200 trained with a smaller batch size (128b_200e) and with the same batch size (200b_200e). By varying the batch size dimension, we defined the different number of samples that are used at training time. Each autoencoder has been implemented by using Keras API [12] with a TensorFlow backend [13]. The attitude assessment has been performed by mean of the following algorithms:

- K-nearest neighbors (KNN) - It is a non-parametric algorithm used for classification and regressions. The class membership of each point is computed from a majority vote of the nearest neighbors of each point: a query point is assigned the data class which has the most representatives within the nearest neighbors of the point. In our case, we used K =3
- Random Decision Forest (RDF) - It is an ensemble learning method that works by building multiple decision trees at training time. The mode of the classes returns to the corresponding class. We set a maximum of 64 trees in the forest and the entropy function to measure the quality of the split.
- Support Vector Regression (RBF) - It is the kernel-based method used for non-linear regression. It defines a kernel matrix computed by applying a pre-defined kernel function to data pairs. We set the kernel function as the standard isotropic Radial Basis Function with Radial Basis Function (RBF). We specified the penalty parameter C to its default value to prevent the model from overfitting.

For these methods, we used the sklearn Python toolbox for Machine Learning [14] in the Google Colaboratory cloud service [15]. The effectiveness of each algorithm is estimated in terms of accuracy (A), precision (P), recall (R), F-Measure (F) and execution time (T). The same metrics are also used to compare the performance of the three algorithms in analyzing the two datasets. The subscript of each metric refers to the analyzed dataset (i.e. U=unimodal and M=multimodal).

IV. RESULTS

The longest interaction session lasted 8 minutes, while the shortest one lasted 2 minutes and a half. The reliability of the conversation with the robot has been affected by different events. Most of the time the robot is not able to understand the request of the user (87%), either because the user is asking some questions that the robot does not know, either because the user speaks with a strong Italian dialect. Due to this misunderstanding, the user attitude changes over time and it was annotated by the human expert during the off-line

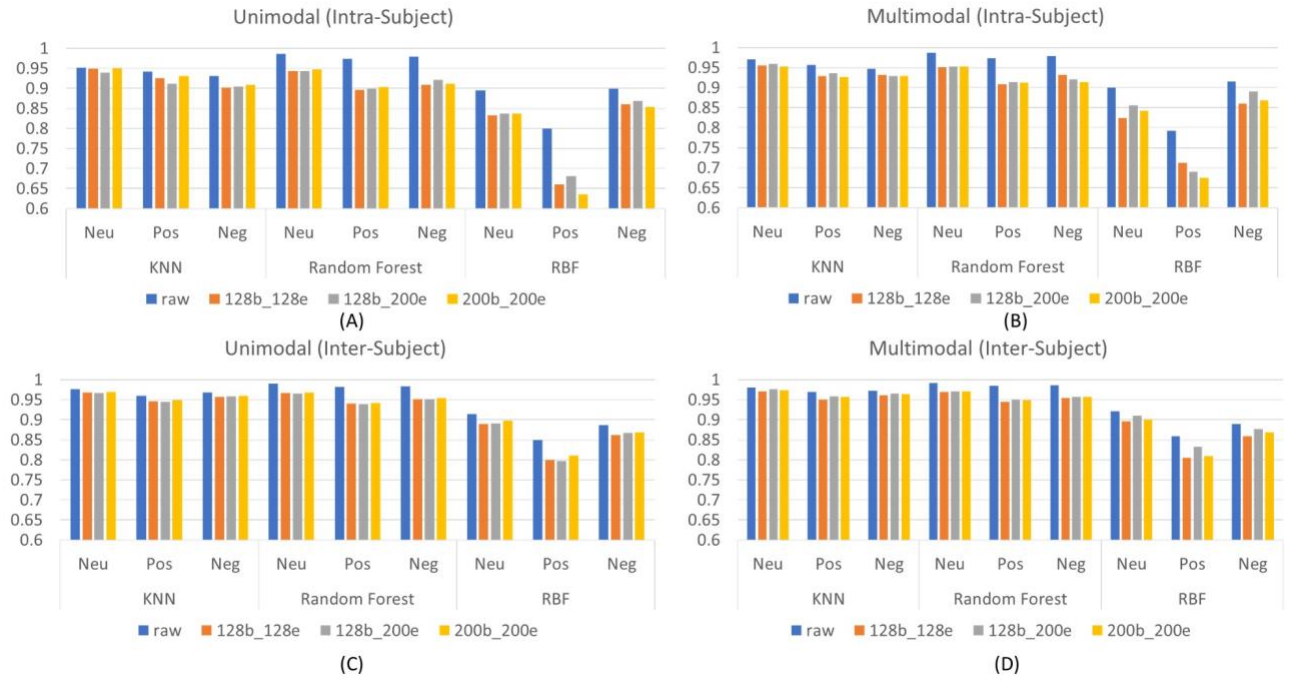

Figure 3. F-Measure performance over the three attitudes of the Unimodal dataset for the intra- (A) and inter-(C) subjects analysis and the multimodal dataset for the intra- (B) and inter-(D) subjects analysis.

process. The annotations report that: the robot answers correctly to the user only the 13% of the cases, the user repeated the questions 15.23% of times and in the remaining times, either the robot does not provide any answer (34.43%), either the robot does not understand the question (34.43%).

It is worth noticing that, one of the main issues of the video and audio recordings concerns missing data, either because the elder is not present in the field of view of the camera either because the caregiver is speaking. In Fig. 2, the diagonal elements show the fraction of data where the open-tools detected the features of interest, whereas the off-diagonals show the fraction of data where the two feature types were extracted simultaneously. The disparity between the fraction of facial landmark detected by OpenPose and the fraction of data detected by OpenFace is related to the use of a different confidence interval ([0,1] for OpenPose, 0-1 for OpenFace). It is worth noticing, that audio features are detected in a small portion (29%) over the datasets.

In the intra-subject validation, the number of samples of the training and testing dataset depends on the number of frames recorded for each user. Even if the percentage of each class is preserved in the 10-CF validation, there are no samples labeled as a negative attitude for user 4 and user 8. These two cases are kept in the inter-subject validation analysis and discarded from the other one. Dealing with the raw representation, RDF algorithm got the higher accuracy (avg. >0.98) for every user both for unimodal and multimodal data. As reported in Table II, the KNN has an average accuracy equal to 0.97 and it is the faster algorithm (T_u=3.88 s, T_m=3.46 s). The performance of RBF algorithm is the worst in terms of accuracy (A_u=0.91, A_m=0.90), F-measure (F_u=0.86, F_m=0.87), precision (P_u=0.90, P_m=0.91), recall (R_u=0.8, R_m=0,77) and execution time (> 11.51 sec) with respect to the other 2 algorithms. In the RBF case, the accuracy of the unimodal data is higher than the accuracy of the multimodal data. Dealing with the auto-encoded representation, the KNN of the 128b_128e case returns the higher accuracy (P_u=0.95, P_m=0.96) in the shortest interval of time (T_u=1.01 sec, T_m=0.85 sec). In the 128b_200e and 200b_200e cases, the KNN algorithm is always the fastest (< 2 s), while in terms of accuracy, both KNN and RDF reach good results (> 0.95) on the unimodal and multimodal datasets. Similarly, the RBF performances are characterized by a lower accuracy (i.e. in the 128b_200e case P_u=0.85, P_m=0.87), lower F-measure (i.e. in the 128b_200e case F_u=0.79, F_m=0.81), lower precision (i.e. in the 128b_200e case P_u=0.84, P_m=0.87) and recall (i.e. in the 128b_200e case R_u=0.73, R_m=0.79) and higher execution time (i.e. in the 128b_200e case T_u=7.5 s, T_m=10.7 s). It is worth noticing that RBF does not assess any negative instance of user 12 both in raw and auto-encoded representation. Fig. 3(A) and Fig. 3(B) show the F-measures obtained in the intra-subject validation for each attitude class.

In the inter-subject validation, the complete dataset is composed of 64530 instances, where 30840 belongs to a neuter attitude, 16590 belongs to a positive attitude and 17100 belongs to a negative attitude. When using the 10-CF validation technique, we ensured that the percentage of samples of each class is preserved. Generally, the performance of the classification with multimodal data is slightly higher than the performance obtained with unimodal data. As shown in Table II, the higher accuracy is achieved by the RDF algorithm for multimodal data with raw representation (A_m=0.987). An accuracy > 0.954 is achieved by KNN and RDF in all the encoded representations. A similar trend is shown also in terms of F-measure (RDF in raw representation F_u=F_m=0.98; RDF in encoder representation F_u=F_m=0.95), as shown in Table II. On the

other hand, the execution time of training and testing of every classification algorithm is extremely high for raw data, especially in the RBF case. The auto-encoder representation drastically decreases the execution time. It is worth noticing that in the KNN and RDF case, the execution time is almost comparable between the different encoder sizes of unimodal and multimodal representation. The comparison of the F-scores obtained in the two validation procedures shows that the neutral and negative attitudes are clearly detected, especially in the multimodal dataset. On the contrary, Fig. 3 highlights the difficulty to assess the positive attitude, especially by the RBF algorithm in the multimodal case.

V. Discussion and Conclusion

The main goal of this work was to investigate whether the available automatic tools can enhance the detection of the elderly attitude. The results show that the visual features extracted by the video make the attitude assessment more robust. Audio information slightly improves the accuracy of the classification. In this work, no enough audio data were belonging to each attitude state to investigate this modality alone.

Regarding the intra-subject analysis, KNN and RDF get high performances in terms of accuracy for raw representation. Acceptable values of accuracy are reached also by the auto-encoded representations. The 128b_128e auto-encoded representation is preferred because drastically reduces the execution time. It is worth noticing that the KNN's execution time in the multimodal dataset is lower than the same in the unimodal dataset. It highlights that the multimodal dataset contains some features which help in the classification. The results obtained in the intra-subject validation shows that assessment can be personalized for each user, like in [6]. To reach it, a dataset composed of a balanced number of instances belonging to each attitude should be available. On the contrary, the inter-subject validation does not suffer from unbalancing results. High performances are achieved by KNN and RDF, as in the intra-class validation case. Comparing the auto-encoded configurations, the auto-encoders of higher size (200) do not improve the classification performances as expected. The overall trend resembles the one of the autoencoder with a smaller size (128). Due to the reported results, the auto-encoded representation with size 128 successfully detects the nonlinearity which may be present in the features set. This evidence suggests the possibility to integrate this representation in a broader architecture.

In the real scenario, autoencoders provide a feasible solution to the online perception of the robot of the user's behavior. This work shows that the compressed information output by the encoder contains the most relevant features detected by the perception system of the robot, reducing the time of the attitude's assessment (see Table II). The flow of information proposed in this work can be integrated into a cognitive architecture to shape the behavior of the robot according to the behavior of the user. The attitude information can be used not only to define what to do but also how the robot should perform the task. The presented work is conducted with a relatively small sample of data. As a preliminary work, we used common machine learning algorithms to evaluate the attitude state of the user. More advanced neural architectures can be introduced once more data are available. We strongly believe that by introducing neural architecture, the robot can automatically assess online what has been performed offline. However, this work shows some limitations which will be addressed as future improvements. The dataset is characterized by a reduced set of users, which leads to an unbalanced quantity of instances belonging to each attitude state. It leads to poor performances in the classification of the positive attitude while achieving good performances in the neutral class. Furthermore, it will be interesting to identify the emotional aspects which are descriptors of the attitude state. Running the classification on a larger set of emotional states could enrich the analysis of the performance and the automatic perception of the technology by the elderly users. One of the future improvements of this work relies on fine graining the list of features that are more representative of the attitude state of the user. This work shows that good performances are achieved by using only visual features. Among that statement, deeper analysis can be carried out to figure out which features highly contribute to the assessment's performance.

References

[1] A. C. Horstmann and N. C. Krämer, "Great Expectations? Relation of Previous Experiences With Social Robots in Real Life or in the Media and Expectancies Based on Qualitative and Quantitative Assessment," *Front. Psychol.*, vol. 10, p. 939, Apr. 2019.

[2] A. Vinciarelli, M. Pantic, and H. Bourlard, "Social signal processing: Survey of an emerging domain," *Image Vis. Comput.*, vol. 27, no. 12, pp. 1743–1759, 2009.

[3] O. Nocentini, L. Fiorini, G. Acerbi, A. Sorrentino, G. Mancioppi, and F. Cavallo, "A survey of behavioural models for social robots," no. May, 2019.

[4] B. De Carolis, S. Ferilli, and G. Palestra, "Simulating empathic behavior in a social assistive robot," *Multimed. Tools Appl.*, 2017.

[5] A. Tanevska, F. Rea, G. Sandini, L. Cañamero, and A. Sciutti, "A Socially Adaptable Framework for Human-Robot Interaction," *arXiv Prepr. arXiv2003.11410*, pp. 1–23, 2020.

[6] O. Rudovic, J. Lee, M. Dai, B. Schuller, and R. W. Picard, "Personalized machine learning for robot perception of affect and engagement in autism therapy," *Sci. Robot.*, 2018.

[7] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *IEEE Trans. Pattern Anal. Mach. Intell.*, 2013.

[8] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science (80-.).*, 2006.

[9] Z. Cao, T. Simon, S. E. Wei, and Y. Sheikh, "Realtime multi-person 2D pose estimation using part affinity fields," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 7291–7299.

[10] T. Baltrusaitis, A. Zadeh, Y. C. Lim, and L. P. Morency, "OpenFace 2.0: Facial behavior analysis toolkit," in *Proceedings - 13th IEEE International Conference on Automatic Face and Gesture Recognition, FG 2018*, 2018.

[11] F. Eyben, F. Weninger, F. Gross, and B. Schuller, "Recent developments in openSMILE, the munich open-source multimedia feature extractor," in *MM 2013 - Proceedings of the 2013 ACM Multimedia Conference*, 2013.

[12] F. Chollet, "Keras Documentation," *Keras.Io*, 2015. .

[13] M. Abadi *et al.*, "TensorFlow: A system for large-scale machine learning," in *Proceedings of the 12th USENIX Symposium on Operating Systems Design and Implementation, OSDI 2016*, 2016.

[14] F. Pedregosa *et al.*, "Scikit-learn: Machine learning in Python," *J. Mach. Learn. Res.*, 2011.

[15] E. Bisong and E. Bisong, "Google Colaboratory," in *Building Machine Learning and Deep Learning Models on Google Cloud Platform*, 2019.