Behavior and Personality Analysis in a nonsocial context Dataset

Dario Dotti, Mirela Popa, Stylianos Asteriadis
Department of Data Science and Knowledge Engineering
Maastricht University, Maastricht, The Netherlands
{dario.dotti, mirela.popa, stelios.asteriadis}@maastrichtuniversity.nl

Abstract

Personality recognition using nonverbal behavioral cues is a challenging task in the Affective Computing field. The majority of existing methods investigate personality assessment in social contexts, such as crowded places or social events, but ignore the role of behaviors as well as personality in nonsocial situations (i.e. during individual activities). In this paper we introduce a novel dataset for behavior understanding and personality recognition in a nonsocial context. Forty-six participants were recorded in an unconstrained indoor space, related to a smart home environment, performing six tasks resembling Activities of Daily Living (ADL). During the experiment, personality scores were collected using self-assessment questionnaires. Furthermore, a temporal framework using a Long-Short Term Memory (LSTM) network is proposed to map nonverbal behavioral features to participants' personality labels. Our experiments showed that nonverbal behaviors are important predictors of personality, confirming theories from the personality psychology field.

1. Introduction

Personality recognition using behavioral observations is a challenging task due to psychological, as well as technical modeling reasons [6]. First of all, underlying human mechanisms for emotion and personality understanding are still mostly obscure to the psychology society. Additionally, human judgment regarding personality evaluation of others is often too unstable, due to many possible interpretations of human expressive power [12]. Research in Affective Computing has shown big improvements over the last years, where verbal and nonverbal behavioral cues have been studied for a variety of applications, such as Human-Computer Interaction (HCI) and Ambient Assisted Living (AAL) [28].

Modeling human behavioral cues requires a deep understanding of several components like facial expressions, gaze, hand gestures, body postures and conversation dynamics. Although facial analysis has provided interesting applications such as automatic job screening [22], it limits the user to be in a frontal position with respect to the computer and in a controlled environment. On the other hand, nonverbal behaviors of human body have been shown to be robust, as well as an important predictor for personality [3].

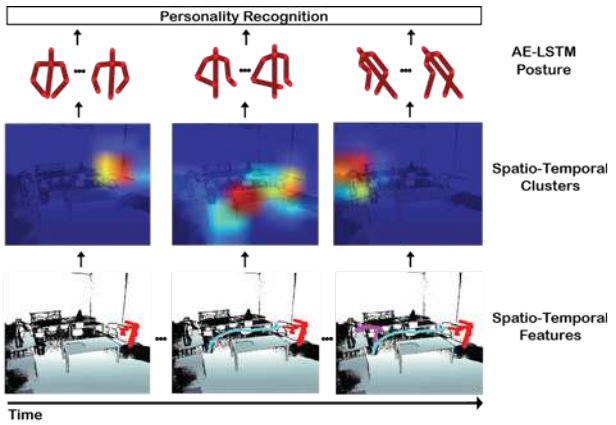


Figure 1. Architecture of the proposed system. First, spatio-temporal clusters are derived from skeleton motion features and spatial heat-maps. Second, from every cluster, LSTM network is used to map posture sequences to personality recognition.

In light of the fact that individuals' interactions with others are shaped by their personality [6], nonverbal behavioral cues have been widely studied in social situations. For example, as shown by [2], extrovert personalities tend to engage in more face-to-face positions during conversations, and in [4], shy personalities tend to avoid walking too close to their neighbours. Nevertheless, psychological research showed that psychological components such as social pressure, may affect natural personality displays in social contexts [7]. Furthermore, for applications in AAL where, for example, a considerable number of people live alone, and reporting feelings of loneliness or social isolation, there is a concrete need for systems able to understand behaviors and personality in nonsocial contexts.

Therefore, this paper introduces a novel dataset, where forty-six subjects perform six tasks divided in three ADL

types: searching for objects, problem solving activities, and daily routine activities. To the authors knowledge, this is the first dataset that provides sensory data, skeleton tracking information and self-assessed personality labels [25], for behavior and personality modeling in an unconstrained indoor environment. Additionally, we propose a nonverbal behavior analysis based on skeleton motion features using Histograms of Oriented Tracklets (HOT) [20], spatial heat-maps, as well as body posture features extracted in an unsupervised way using Autoencoders [19]. Moreover, as behaviors have a dynamic nature, temporal sequences are investigated in a Long-short term Memory (LSTM) framework, for personality recognition (Fig. 1).

The contributions of this paper are as follows: Firstly, we introduce a novel dataset for personality assessment, where forty-six subjects were recorded performing six activities. The recorded behavior was spontaneous, as the participants did not receive beforehand any indication on how to perform the tasks or about the goal of the study. Secondly, we propose a novel framework for personality recognition, that encodes spatio-temporal features, as well as body postures in an unsupervised way, and models behavior dynamics using the LSTM network. Thirdly, by clustering the participants big five personality scores, we obtain higher dependencies between traits, achieving results that are consistent with the psychological theory proposed in [5].

2. Related Work

Personality Datasets. Personality computing has received increasing attention in the computer vision community, and consequently, several datasets have been proposed. In [2], a dataset for behavior and personality analysis using Free-standing Conversation Group (FCG) is proposed. The authors used a multimodal approach, where audio, video, infrared, Bluetooth and accelerometer data were recorded in an unconstrained social environment for behavior understanding. Annotations for body detection and personality were provided, and personality scores were correlated with the analysis of social interactions and group formation. Instead, we would like to highlight that our dataset analyzes behaviors of individuals that are alone in a room, concentrating on the display of personality during execution of different tasks. Studies linking face analysis, audio information and personality have shown progress in the last years. First Impressions challenge dataset was proposed in [22]. The goal of the data was to automatically evaluate the personality of subjects for a job screening application. Findings were evaluated in a quantitative way, as no psychological theories were linked to them. Therefore, in this paper, we highlight the need of interdisciplinary research, between psychology, computer vision and affective computing, for enhancing the ability to understand and automatically recognize human personality on a novel benchmark dataset.

Nonverbal behavioral features. *There is something in the nature of individuals that leads observers to attribute certain characteristics to them* [3]. Since this pioneering work, nonverbal behaviors have been studied extensively in the fields of psychology and Affective Computing. In [15], body movements of public speakers were analyzed and correlated to personality scores. The obtained results showed that upper body motion (e.g. extracted from head, torso and hands) was a reliable predictor for openness and neuroticism traits. During social interactions, authors in [29] used proxemics and visual attention features to predict *Extraversion* and *Neuroticism* personality traits. In [23], a set of motion and proximity features was extracted for extraversion prediction during Human-Robot Interaction, and similarly, [17] predicts extraversion scores using attention features during a work meeting. In these works, the five personality traits were treated independently, while in the current study, we make use of a clustering technique to find higher dependencies between the five personality traits, confirming the psychological theory proposed in [5].

Posture dynamics using skeleton data. Since the distribution of low-cost depth sensors with skeleton estimation algorithms [27], and the recent success of deep neural network systems like Convolutional Neural Networks (CNNs) [16], joints dynamics have been studied for a variety of tasks like action recognition [9] and posture learning [18]. CNN capabilities are optimized for feature extraction and learning using images, but since the skeleton data has a different structure [10], approaches have been proposed to overcome this issue. In [9], authors represent skeleton sequences as images, to maintain the skeleton structure, joints of each body part are concatenated by their physical connections. Finally CNN networks learn the spatial temporal sequences for action recognition. However, our approach consists of encoding the joints spatial relations using an autoencoder framework, while the temporal information is modeled using an LSTM network. Our proposed posture representation is efficient and simplified, by removing background noise existent in images, while being less demanding in terms of processing power, computation time and amount of training samples needed for CNN models.

3. Behavior and Personality in a nonsocial context Dataset

While human behavior has been largely studied in social environments [2] and crowded places [1], few efforts have been made towards the relation between human behavior and personality in nonsocial situations, e.g. performing individual tasks. In order to provide a new benchmark to further study the relation between human behavior and personality recognition, in this paper we release the Behavior

and Personality in a nonsocial context Dataset. [1]

3.1. Experimental Design

The experimental design was inspired by both ADL datasets [21], as well as problem-solving based psychological tests. To create an unconstrained environment, no time limit nor know-how was given to complete the proposed six tasks. To elicit real personality manifestations in the most unobtrusive way, problem solving level was varied every two tasks. The experimental room was furnished with tables, chairs, a tea corner with a water kettle, and two office cabinets, having each drawers filled with many different objects. All around the room, boxes and cases also containing objects, were spread to challenge our subjects for the completion of the experimental tasks.

Tasks were organized in three groups called: "daily-routine activities" with low problem solving difficulty, "searching-activities" with medium problem solving difficulty, and finally, "problem-solving activities" with high difficulty. Once every task was completed, participants were asked to exit the room to record the task completion through magnetic sensors positioned on the door. In the searching-activities, participants were requested to: 1) find keys of two cabinets present in the room, and to 2) find an item, the experimenter hid beforehand. These two activities resemble daily activities of searching for items at home. In the problem-solving activities, participants had to: 3) search for an item that was not in the room, and to 4) memorize the content in all drawers of the two cabinets. Since in the first group of tasks, the participants found the requested object, no one expected that in task 3 there would be no object. After searching the room thoroughly, most of the participants felt confused and started to wander around the room without an evident goal. In the daily routine activities, we asked the subjects to: 5) sit at the table to complete two questionnaires, and to 6) make tea and eat cookies. These last activities are inspired from computer vision datasets and can be used by the research community to further study the relationship between activity and personality.

3.2. Video Data

The data was recorded using Kinect SDK 2.0 released by Microsoft. The SDK skeleton tracking functionality detects and tracks 20 joints on the human skeleton at around 30 frames per second. The coverage range in which the tracking algorithm provides reliable results is from 0.5 meter to 5 meters, while outside this range, skeleton information was less reliable and was removed. The Kinect sensor was placed at around two meters distance from the ground to be robust from occlusions. Forty-six young adults were recorded for a total of 550 minutes, 16500 frames with annotated skeleton joints information were stored and depth images were saved at the rate of 6 frames per second. For privacy reasons, we decided to record only the depth images and not the color images (Fig. 2).

[1]Dataset is available at `https://project.dke.maastrichtuniversity.nl/personality/`

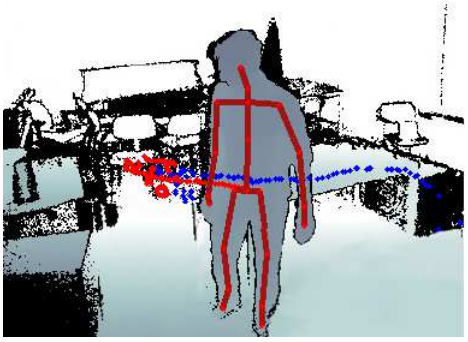


Figure 2. Example of the data released in this dataset, containing the 3D joint coordinates, the user trajectory in each task, as well as the depth image.

3.3. Personality Data

In order to collect personality scores from each subject in a unobtrusive way, we used the short version of the Big Five Inventory (BFI-10), introduced in [25]. This short version measures the Big-Five traits, namely, Extraversion, Agreeableness, Conscientiousness, Neuroticism and Openness, using 10 questions (i.e. two questions per personality trait) and can be completed in less than one minute. Even though shorter than the BFI-44 [14], it was shown to provide reliable scores because it was built by preserving the questions that best correlate with the results of the original inventory. In this study, we followed a first-person (self-assessment) strategy, where participants were told that all questionnaires would remain anonymous in order to avoid biased answers.

4. Method

In this paper, we investigate nonverbal behavioral cues for personality recognition on a novel dataset. Due to the position of the camera, upper body joints are the most robust to occlusion and noise, therefore, in our feature extraction phase, we consider only eight joints: head, spine, left-right shoulders, elbows and wrists. For each frame, behaviors are represented in terms of joints local spatial relation (posture), motion, as well as global spatial location (heat-map of the room). In this section, the aforementioned features are explained in details.

4.1. Unsupervised Posture Representation

In this paper, we introduce a novel approach to learn upper body posture representation using autoencoders [19].

Our approach consists of two parts: posture extraction and posture learning.

Skeleton-based representation has shown promising results in encoding the spatial-temporal relation among joints for action recognition using neural networks architecture [9]. Therefore, we propose a new descriptor which is optimized for an autoencoder framework and it aims to learn in an unsupervised way the skeleton joints local relation through posture.

For every frame, we build a new binary image of size $s \times s$ around the upper body skeleton data, where the pixels corresponding to the eight joints of interest are set to a value equal to one, and the other pixels are considered as background, with a value equal to zero. The image size is selected to account for all possible situations in which the joints could appear (e.g. when the arms are wide open the overall posture size is bigger than when the arms are closed). Single skeleton coordinates x, y are too sparse to be learned in an efficient way, and moreover, the pose information conveyed is limited. Hence, following their natural physical connections (i.e. left shoulder and right shoulder), related joints are connected by a line with a value equal to one. Even though the 3D skeleton coordinates are converted into a 2D descriptor, the z coordinate is not lost. During the posture extraction, the coordinates are not normalized, nor centralized, resulting in a raw skeleton descriptor that embeds the z coordinate (e.g. distance to the camera). Frames, where the skeleton is far away from the Kinect sensor (high value of z), will contain a body posture with a smaller size than the ones in which the skeleton is closer to the sensor (Fig 3). The advantages of our descriptor are the following: 1) we preserve the local spatial relation between joints for posture learning, and 2) we reduce the learning problem complexity by using a binary image, where the desired information is set to a value equal to one, and the background noise is set to zero.

Encouraged by the impressive results of autoencoders in image reconstruction, a deep autoencoder is trained to minimize the input reconstruction error [2]. For each autoencoder layer l_a, the encoder f^{l_a} and decoder g^{l_a} functions are designed to reconstruct the input data X, represented as a vectorized set of input features $X_i = [x_1, \ldots, x_n]^T \in R^n$, as good as possible in an unsupervised way. Therefore, given input data X_i, the encoding step is obtained using the function f^{l_a}, while the mid-level representation is denoted by $\alpha^{l_a}(i) = f^{l_a}(W_1^{l_a} \cdot \alpha^{l_a-1}(i) + b^{l_a})$ and the decoding step is captured by the function g^{l_a}. On the first layer we consider $\alpha^0(i) = X_i$ and on the following layers the input will be represented by the projected data in the hidden units space learned on the previous layer, so by α^{l_a-1}. The reconstruction result is denoted by $\hat{X}_i = g^{l_a}(W_2^{l_a} \cdot \alpha^{l_a}(i) + c^{l_a})$, while on the following layers the reconstruction will be represented by $\hat{\alpha}^{l_a-1}$. $\{W_1^{l_a}, W_2^{l_a}\}$ are the weight matrices and $\{b^{l_a}, c^{l_a}\}$ are the encoding and decoding bias parameters on each layer l_a of the autoencoder framework. The optimization goal is to minimize the error between the input data X_i and the reconstructed data $\hat{X}_i$, using stochastic gradient descent with adaptive learning rate and cross-entropy (CE) cost function defined as:

$$CE(X, \hat{X}) = -\sum_{i=1}^{n} x_i \log(\hat{x}_i) + \sum_{j=1}^{n_{l_1}} \lambda \|\alpha^{l_1}(x_j)\|_1 \quad (1)$$

In Fig. 3, we start by visually inspecting the reconstructed images using the deep autoencoder (with $l_a = 2$) learned weights with $n_{l1} = 900$ hidden units in the first layer and $n_{l2} = 225$ units in the second layer, values obtained for the optimal value of the cost function in Eq. 1. Different postures are clearly visible, where the 2D skeleton information x, y embeds spatial relation between the joints. Moreover, our learned representation shows to be robust to skeleton size changes, embedding the third coordinate z.

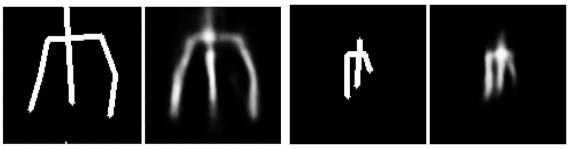


Figure 3. Raw posture descriptors and corresponding reconstructions. The autoencoder weights learn the 2D joints relation, while depth information is retained through the relative size of the skeleton.

4.2. Motion Features: Skeleton HOT

In [20], authors showed that analyzing the spatio-temporal descriptors called tracklets, could improve the recognition of human motion. A tracklet indicates the movement of a subject, for a short period of time. In our unconstrained experimental design, tracklets are considered more robust, as well as, they can be more meaningful than global trajectories. Indeed, due to the fact that ultimately our subjects are completing the same tasks in a limited scenario, subjects' global trajectories risk to coincide. Moreover, it has been shown that thin slices of motion cues are enough for personality prediction [3]. Hence, an adaptation of the Histograms of Oriented Tracklets [20],[8], is used to encode the magnitude and orientation of each upper-body joint in a time window t_τ. For every frame i, we compute the magnitude and orientation of each upper body joint j with respect to the previous frame, and the obtained values are quantized in $M = 3$ bins for magnitude and $O = 8$ bins for orientation. Every histogram descriptor accumulates values within a time window $t_\tau = 1$ second. Note that

[2] For our experiments we used NVIDIA Titan X GPUs.

since body joints movements are highly correlated, we subtract the magnitude and orientation value of the head joint from all the other upper body joints.

4.3. Spatial heat-map

In every scenario, behaviors are correlated with areas where they are performed, for example, the activity of making tea takes place at the tea corner, whereas walking behaviors happens in the walking area. Therefore, we propose a heat-map descriptor to correlate spatial information of the scene with the degree of occupancy in each area. Firstly, the video scene is divided into 3D nonoverlapping patches, where every cube is of size $h \times w \times d$, namely height, width and depth. Secondly, the heat-map descriptor is built by counting the occurrences of the upper-body skeleton joint coordinates inside each cube for every time window t_τ. In a time window where $t_\tau = 1$ second, our heat-map descriptor indicates where the subject is in the scene, and additionally, provides insights about the observed behaviors. If the subject is moving, trajectory points can be found in more than one patch, whereas if the subject is stationary, points can be found in only one patch.

4.4. Posture dynamic modeling and personality recognition using LSTM

Although the features described above have been shown to be important indicators of personality [15], [23], the temporal nature of human behaviors also plays an important role in personality recognition. For this reason, we propose to use an LSTM network to learn behavior dynamics, and, by adding a classification layer on top of the LSTM output, recognition of the participants' personality type is performed.

LSTM networks have shown good accuracy for movement learning, however, in cluttered and noisy scenarios, a single LSTM network lacks learning capacity [1]. Hence, inspired by [11], cluster analysis is performed on the spatial heat-map and skeleton HOT information, to reduce motion ambiguities given by the unconstrained experiment scenario. Finally, LSTM networks are trained on every cluster data, for posture sequence learning and personality recognition (our system architecture is shown in Fig. 1).

In this work we chose to employ the Gaussian Mixture Models (GMM) technique, due to its ability to maximize the component posterior probability given the data. GMM clustering [30], is applied on the spatial heat-map (Section 4.3), and the skeleton HOT information (Section 4.2). Our goal is to find a set of clusters $c_1, .., c_k$ that defines a clear separation between behavioral patterns, for example, searching for an object produces different spatio-temporal information than the activity of making tea, as they are happening in different regions of the scene and they are characterized by different motion magnitudes. Given the ultimate goal of personality recognition, we believe that splitting the posture data in different behavioral patterns, encourages the LSTM in learning to map joints spatio-temporal relation to personality labels, by reducing the behavioral variability.

Posture features are extracted from each generated cluster as displayed in Fig. 1, and are encoded for every frame using the deep autoencoder technique (Section 4.1). A limitation of our method is that some clusters do not have enough data to encode posture sequences of 30 fps. For this reason, we empirically down-sampled the skeleton sequence to 8 fps, without loss of information. Finally, for each cluster c, a LSTM network is trained to map posture sequences $x_1, .., x_l$ of length $l = 8$ to personality type $y, y \in \{1, \ldots, n_y\}$, where each sequence item x_i is encoded by the deep autoencoder and contains 225 features as described in Section 4.1. Our training objectives are twofold: firstly, we aim to learn the posture dynamics in each spatio-temporal cluster, and secondly, we aim to capture the relations between behavior display and the associated personality label.

In this work, we implemented an LSTM network as in [13], that uses memory cells h^t for each time slot, with input gates i^t, forget gates f^t and output gates o^t, applied on the input node g^t, and as output, a dense layer followed by an element-wise sigmoid activation function, for enabling a multilabel classification. The log loss function (cross entropy) is used at each output (Eq. 1), to learn the true personality label y.

5. Experiment and Results

In this Section, the experiments are explained in detail. Firstly, data analysis on participants' personality scores provides the ground truth label for our personality recognition. Secondly, personality recognition using an LSTM framework is carried out to investigate the relation between low level noverbal behavioral features and personality display.

5.1. Personality Data analysis

Affective Computing is dominated by the dimensional approach of the Five Factor Models [28], where traits are considered independently. However, this approach fails in considering the configuration of these traits within a person. In contrast, numerous studies confirmed the theory proposed in [5], in which all the personality traits can be organized in three major types: resilient, undercontrolled and overcontrolled. Resilient personality type showed below average neuroticism, and intermediate or above average for the rest of the traits, the undercontrolled type usually scores high in neuroticism and extraversion, and finally, the overcontrolled type scores below average on extraversion and above average on neuroticism.

In this study, we propose to follow a data-driven approach, applying a clustering technique on the participants

personality scores, to find a higher convergence of traits. The results of the short BFI version [25] give the score of the five traits on a 1-10 scale, hence, every subject is represented by a vector $v = 1 \times 5$. Hierarchical clustering was applied, and three main clusters were found automatically ($n_y = 3$).

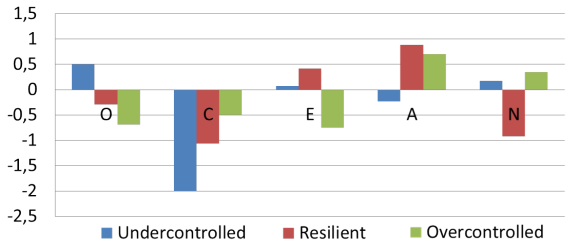


Figure 4. Z-score values of the personality traits in each personality types (undercontrolled, resilient, overcontrolled).

Following the same procedure as in [26], we display the z-score of the three clusters compared to the population mean provided by [24]. In Fig. 4, we show the results which are consistent with the theory proposed in [5], as the resilient personality type has a lower neuroticism score and a higher extraversion score than the population mean. The undercontrolled type exhibits extraversion as well as neuroticism above the population mean, and finally, the overcontrolled personality type showed a low score in extraversion and a high score in neuroticism. The resilient cluster was found to be the most populated with 21 participants, the overcontrolled cluster contains 15 participants, and finally, the undercontrolled cluster contains 10 participants. These findings provide a new way of labelling personality, and will be used as ground truth for our personality recognition experiments. The novelty of the approach is given by the fact that our model is able to learn directly different combinations of the big five traits using only three labels.

5.2. Spatio-Temporal Clustering of nonverbal behaviors

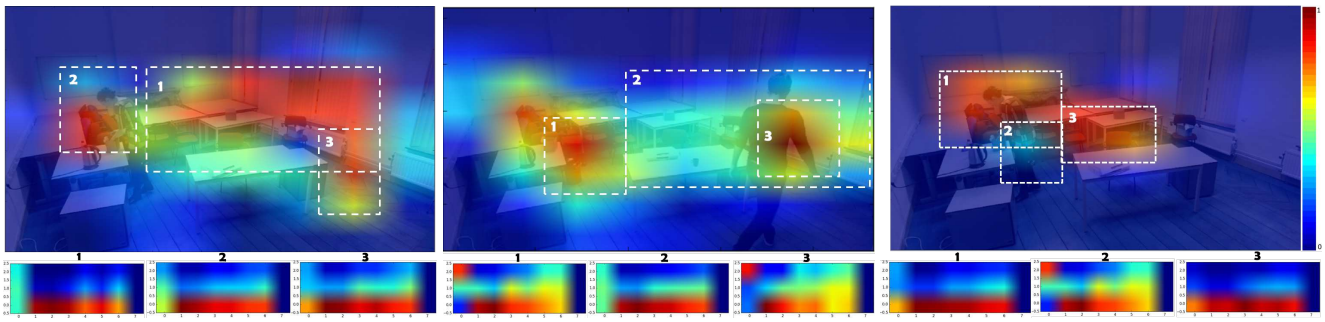
In the proposed study, we aim to map nonverbal behavioral features to personality labels for personality recognition. The proposed benchmark is very challenging due to its unconstrained structure, where participants could adopt any strategy to complete the tasks. Therefore, in order to reduce motion ambiguities, for each activity type explained in Section 3.1, we aim to obtain spatio-temporal clusters representing different behavioral patterns.

Skeleton motion features (Section 4.2) and the heat-map descriptors (Section 4.3) are extracted at an interval of $t_\tau = 1$ second(s) and GMM clustering is applied to find k number of clusters. The optimization of the intra-clusters variance with respect to the total variance is used to select the correct number of clusters k. For the searching-activity task $k^1 = 17$ clusters were found, for the problem-solving activity task $k^2 = 19$ and finally, for the daily-routine activity tasks $k^3 = 16$ clusters were found.

In Fig. 5, we show the top three most populated clusters in each activity type. In the first row, the spatial heat-map information is displayed, where pixel values range from dark blue, which shows that the area is not visited, to dark red, which defines areas that are visited the most. The second row shows the mean values of the skeleton motion information in the respective clusters. Only the $M = 3$ magnitude bins (y-axis) are considered for each of the $j = 8$ upper-body joints plotted in the x-axis with the following order: head, left shoulder, right shoulder, left elbow, right elbow, left wrist, right wrist and spine. Pixel values range from dark blue, which shows that a bin is not populated, to dark red, which defines a population with a high density. As displayed in Fig. 5, the spatio-temporal clusters provide a salience map of the scene for each activity type. In the searching-activity tasks, Fig. 5(a), participants walked around looking for items, covering many parts of the scene. This behavior is reflected in the motion information, where all three clusters, report movements of almost all the joints. In particular, motion information is very informative for defining meaningful sub-behaviors of searching activities. For example, cluster number one contains high movement of the head (joint number zero), but slow movement of the other joints, which characterize a walking behavior. On the contrary, in cluster number two and three, motion from the other joints is detected, showing that participants were exploring the content of the scene in order to complete the tasks. Fig. 5(a) and Fig. 5(b) display similar behavioral patterns, where participants were challenged to fulfill the tasks. On the other hand, daily-routine activities (Fig. 5(c)), are characterized by a different spatio-temporal salience map. Participants were mainly concentrated in the areas of the table (filling questionnaire task) and the tea corner (making tea activity). In cluster one, movements of the head and arms (elbow and wrist joints) indicate that participants were performing the activity of making tea, whereas in cluster number three, where participants were sitting at the table, the reported joint movement is little. Our cluster analysis provided spatio-temporal separations between behavioral patterns in an unsupervised manner (e.g. searching behaviors versus making tea behaviors). Hence, for each cluster, a LSTM network is trained using posture features introduced in Section 4.4.

5.3. Personality Recognition

In order to examine the strength of our proposed framework, in this section, two personality recognition experiments are reported. We compare the performances of the LSTM combined with clustering ($LSTM_{cl}$), with a basic LSTM framework ($LSTM$), where, for each activity type,



(a) (b) (c)

Figure 5. Clustering analysis on the three activity types: (a) searching-activities (b) problem-solving activities and (c) daily-routine activities.

an LSTM network is trained for personality recognition.

In the first experiment "cross-participant", the mapping between posture sequences and personality labels is evaluated in every cluster, where we use 90% of the data across participants for training and 10% for testing in a 5-fold cross validation. The final accuracy ($LSTM_{cl}$) variant, is obtained computing the mean of the all clusters accuracy. Whereas, for the basic LSTM, ($LSTM$), the final accuracy is the average accuracy obtained on the test sample for each activity type. In the second experiment "per-participant", we test our system per participant (e.g. leave-one-out scheme), where we use all the participants data except one to train our two LSTM variants, and the left-one-out for testing.

In Table 1, we report the classification F1 score as well as the cross-entropy error (CE) across participants. Note, that due to the fact that LSTMs are trained using a cross-entropy loss function, we report *CE* as the probability that the predicted label $\hat{y}$ is equal to the true label y (Eq. 1). Results highlight that even a general LSTM framework can successfully map the proposed nonverbal behavioral features to personality display. Furthermore, it is evident that separating behavioral patterns using a clustering technique, reduces the ambiguity of the posture sequences and improves the personality recognition in each activity type. The best accuracy result is obtained in the daily-routine type, where participants were asked to perform daily activities, experiencing less pressure and a low level of challenge. This set-up allowed them to create more relaxed and smooth movements, that were better captured by our autoencoder-LSTM framework.

In the AAL context, our framework should be able to recognize the observed personality in real time, given participants' behaviors. In this experiment, we train the two LSTM variants on all the participants except one, and use the test participant p for the evaluation. This process is repeated for all the participants. To overcome the imbalanced personality labels explained in Section 5.1, we randomly

LSTMs	**A1**		**A2**		**A3**	
	f1	CE	f1	CE	f1	CE
$LSTM_{cl}$	**0.615**	**1.298**	**0.6404**	**1.287**	**0.7395**	**0.8765**
$LSTM$	0.5263	1.697	0.5872	1.425	0.7269	0.9342

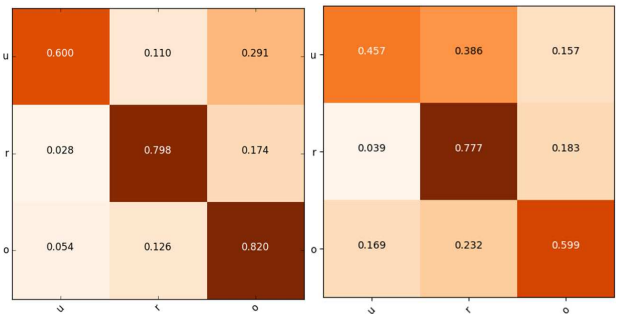
Table 1. F1 accuracy and cross-entropy error (CE) for personality recognition cross-participant experiment, where A1=searching activities, A2=problem-solving activities and A3=daily-routine activities.

LSTMs	**A1**		**A2**		**A3**	
	Recall	CE	Recall	CE	Recall	CE
$LSTM_{cl}$	**0.5148**	**1.455**	**0.5333**	**1.403**	**0.6116**	**1.3897**
$LSTM$	0.4745	1.852	0.5	1.4918	0.5134	1.5109

Table 2. Mean Recall accuracy and cross-entropy error (CE) for personality recognition per-participant experiment, in the three activity types. A1=searching activities, A2=problem-solving activities and A3=daily-routine activities.

under-sample the majority classes, obtaining a dataset with ten participants per class. Therefore, for every participant p in the dataset, we classify all the corresponding posture sequences and we report the mean recall accuracy score, as well as the CE in Table 2. The recall accuracy was chosen in this experiment, because every sample of the test participant has a fix personality label, making the precision accuracy always equal to 1, and therefore biasing the f1 score. The obtained results are in line with the system performance showed in the previous experiment, where the proposed clustering approach improved the basic LSTM in all the activity types, and the best accuracy was obtained in the daily-routine activities.

Finally, to further investigate the personality recognition performance of our system, in Fig. 6 we show the confusion matrix of the daily-routine activity type from both experiments. Overall, the resilient (r) and overcontrolled classes (o) obtained the best results, showing that our LSTM framework could learn to distinguish their different configurations of personality traits. In this sense, resilient and



(a) (b)

Figure 6. Confusion matrix of the personality recognition for daily-routine activities: (a) cross-participants experiment (b) per-participant experiment. Labels are: u: undercontrolled, r: resilient, o: overcontrolled.

overcontrolled classes contain significant differences in the *Extraversion* and *Neuroticism* traits, suggesting that these traits are the most relevant to our experiment.

6. Discussion

To understand which discriminative patterns our LSTM network learns for each personality type, we display in Fig. 7 the sequences that obtained the highest confidence during the recognition phase. In particular, we overlap the top 2 sequences of length $l = 8$, for each personality type during the daily-routine activities. It is observed that the resilient personality (second row), presents significant difference in the movement of the arms with respect to the other personality types. Given its traits configuration (i.e. low Neuroticism and high Extraversion), these sequences may represent relaxed and talkative personality attributes, showing no apprehension towards completing the tasks. On the other hand, the undercontrolled (first row) as well as the overcontrolled (third row) personalities, present sequences with stiffer postures. More specifically, the overcontrolled type, displays arms linked and arched back, as the skeleton is in the act of searching for objects. This nonverbal behavior may represent stress to complete the tasks, as well as no interests in social contact with the experimenter, supporting its traits configuration (i.e. high Neuroticism and low Extraversion).

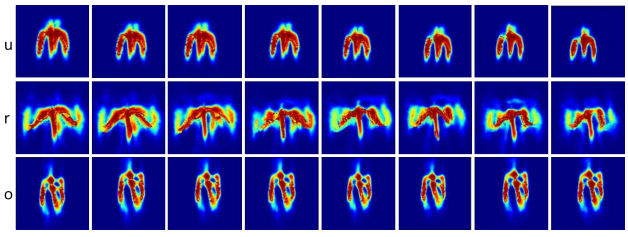


Figure 7. Visualization of the sequences that obtained the highest recognition confidence for each personality type, the labels are: u:undercontrolled, r:resilient, o:overcontrolled.

7. Conclusion

In this paper, we introduced a novel dataset for behavior understanding and personality recognition, recorded in an unconstrained indoor scenario, related to a smart home environment. Forty-six participants performed six tasks belonging to three daily activity types: searching, problem-solving and daily-routine activities. To the best of the authors' knowledge, this is the first dataset that provides depth data, skeleton tracking information for individual behavior analysis and personality labels. Furthermore, we employed an LSTM framework to map nonverbal behavioral features to personality labels. The effectiveness of the proposed framework and the validity of the dataset was demonstrated by two personality recognition experiments, providing new insights regarding the relation between nonverbal behavioral cues and personality types.

8. Acknowledgement

This work has been funded by the European Unions Horizon 2020 Research and Innovation Programme under Grant Agreement N 690090 (ICT4Life project). Additionally, we would like to thank all the participants in our experiment, who were students as well as employees from the University of Maastricht (The Netherlands).

References

[1] A. Alahi, K. Goel, V. Ramanathan, A. Robicquet, L. Fei-Fei, and S. Savarese. Social lstm: Human trajectory prediction in crowded spaces. In *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, pages 961–971, 2016.

[2] X. Alameda-Pineda, J. Staiano, R. Subramanian, L. Batrinca, E. Ricci, B. Lepri, O. Lanz, and N. Sebe. Salsa: A novel dataset for multimodal group behavior analysis. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 38(8):1707–1720, 2016.

[3] N. Ambady and R. Rosenthal. Thin slices of expressive behavior as predictors of interpersonal consequences: A meta-analysis. *Psychological bulletin*, 111(2):256, 1992.

[4] A. Bera, T. Randhavane, and D. Manocha. Aggressive, tense, or shy? identifying personality traits from crowd videos. In *Proc. of the Twenty-Sixth Int. Joint Conf. on Artificial Intelligence, IJCAI-17*, pages 112–118, 2017.

[5] J. H. Block and J. Block. The role of ego-control and ego-resiliency in the organization of behavior. In *Development of cognition, affect, and social relations: The Minnesota Symposia on child psychology*, volume 13, pages 39–101, 1980.

[6] O. Celiktutan, E. Sariyanidi, and H. Gunes. Computational analysis of affect, personality, and engagement in human–robot interactions. 2018.

[7] P. M. Cole. Children's spontaneous control of facial expression. *Child development*, pages 1309–1321, 1986.

[8] D. Dotti, M. Popa, and S. Asteriadis. Unsupervised discovery of normal and abnormal activity patterns in indoor and

outdoor environments. In *VISIGRAPP (5: VISAPP)*, pages 210–217, 2017.

[9] Y. Du, Y. Fu, and L. Wang. Skeleton based action recognition with convolutional neural network. In *Pattern Recognition (ACPR), 2015 3rd IAPR Asian Conference on*, pages 579–583. IEEE, 2015.

[10] G. Ercolano, D. Riccio, and S. Rossi. Two deep approaches for adl recognition: A multi-scale lstm and a cnn-lstm with a 3d matrix skeleton representation. In *Robot and Human Interactive Communication (RO-MAN), 2017 26th IEEE Int. Symposium on*, pages 877–882. IEEE, 2017.

[11] T. Fernando, S. Denman, S. Sridharan, and C. Fookes. Soft+ hardwired attention: An lstm framework for human trajectory prediction and abnormal event detection. *arXiv preprint arXiv:1702.05552*, 2017.

[12] H. Gunes, C. Shan, S. Chen, and Y. Tian. Bodily expression for automatic affect recognition. *Emotion recognition: A pattern analysis approach*, pages 343–377, 2015.

[13] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[14] O. P. John, E. Donahue, and R. Kentle. the big five. *Factor Taxonomy: Dimensions of Personality in the Natural Language and in Questionnaires.In Handbook of Personality: Theory and Research, ed. Lawrence A. Pervin and Oliver P. John*, pages 66–100, 1990.

[15] M. Koppensteiner. Motion cues that make an impression: Predicting perceived personality by minimal motion information. *Journal of experimental social psychology*, 49(6):1137–1143, 2013.

[16] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[17] B. Lepri, R. Subramanian, K. Kalimeri, J. Staiano, F. Pianesi, and N. Sebe. Connecting meeting behavior with extraversiona systematic study. *IEEE Trans. on Affective Computing*, 3(4):443–455, 2012.

[18] J. Liu, N. Akhtar, and A. Mian. Skepxels: Spatio-temporal image representation of human skeleton joints for action recognition. *arXiv preprint arXiv:1711.05941*, 2017.

[19] C. P. MarcAurelio Ranzato, S. Chopra, and Y. LeCun. Efficient learning of sparse representations with an energy-based model. *Advances in neural information processing systems*, pages 1137–1144, 2007.

[20] H. Mousavi, M. Nabi, H. K. Galoogahi, A. Perina, and V. Murino. Abnormality detection with improved histogram of oriented tracklets. In *Int. Conf. on Image Analysis and Processing*, pages 722–732. Springer, 2015.

[21] B. Ni, G. Wang, and P. Moulin. Rgbd-hudaact: A color-depth video database for human daily activity recognition. In *Consumer Depth Cameras for Computer Vision*, pages 193–208. Springer, 2013.

[22] V. Ponce-López, B. Chen, M. Oliu, C. Corneanu, A. Clapés, I. Guyon, X. Baró, H. J. Escalante, and S. Escalera. Chalearn lap 2016: First round challenge on first impressions-dataset and results. In *European Conf. on Computer Vision*, pages 400–418. Springer, 2016.

[23] F. Rahbar, S. M. Anzalone, G. Varni, E. Zibetti, S. Ivaldi, and M. Chetouani. Predicting extraversion from non-verbal features during a face-to-face human-robot interaction. In *Int. Conf. on Social Robotics*, pages 543–553. Springer, 2015.

[24] B. Rammstedt. The 10-item big five inventory. *European Journal of Psychological Assessment*, 23(3):193–201, 2007.

[25] B. Rammstedt and O. P. John. Measuring personality in one minute or less: A 10-item short version of the big five inventory in english and german. *Journal of research in Personality*, 41(1):203–212, 2007.

[26] F. A. Sava and R. I. Popa. Personality types based on the big five model. a cluster analysis over the romanian population. *Cognitie, Creier, Comportament/Cognition, Brain, Behavior*, 15(3), 2011.

[27] J. Shotton, A. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman, and A. Blake. Real-time human pose recognition in parts from single depth images. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conf. on*, pages 1297–1304. Ieee, 2011.

[28] A. Vinciarelli and G. Mohammadi. A survey of personality computing. *IEEE Trans. on Affective Computing*, 5(3):273–291, 2014.

[29] G. Zen, B. Lepri, E. Ricci, and O. Lanz. Space speaks: towards socially and personality aware visual surveillance. In *Proc. of the 1st ACM Int. Workshop on Multimodal Pervasive Video Analysis*, pages 37–42. ACM, 2010.

[30] Z. Zivkovic. Improved adaptive gaussian mixture model for background subtraction. In *Proc. of the 17th Int. Conf. on Pattern Recognition (ICPR 2004).*, volume 2, pages 28–31. IEEE, 2004.