# *Comfortability* Recognition from Visual Non-verbal Cues

Maria Elena Lechuga Redondo
Italian Institute of Technology
Genova, Italy
maria.lechuga@iit.it

Alessandra Sciutti
Italian Institute of Technology
Genova, Italy
alessandra.sciutti@iit.it

Francesco Rea
Italian Institute of Technology
Genova, Italy
francesco.rea@iit.it

Radoslaw Niewiadomski
University of Trento
Rovereto, Italy
r.niewiadomski@unitn.it

## ABSTRACT

As social agents, we experience situations in which sometimes we enjoy being involved and others where we desire to withdraw from. Being aware of others' "comfort towards the interaction" help us enhance our communications, thus this becomes a fundamental skill for any interactive agent (either a robot or an Embodied Conversational Agent (ECA)). For this reason, the current paper considers *Comfortability*, the internal state that focuses on the person's desire to maintain or withdraw from an interaction, exploring whether it is possible to recognize it from human non-verbal behaviour. To this aim, videos collected during real Human-Robot Interactions (HRI) were segmented, manually annotated and used to train four standard classifiers. Concretely, different combinations of various facial and upper-body movements (i.e., *Action Units, Head Pose, Upper-body Pose* and *Gaze*) were fed to the following feature-based Machine Learning (ML) algorithms: *Naive Bayes, Neural Networks, Random Forest* and *Support Vector Machines*. The results indicate that the best model, obtaining a 75% recognition accuracy, is trained with all the aforementioned cues together and based on *Random Forest*. These findings indicate, for the first time, that *Comfortability* can be automatically recognized, paving the way to its future integration into interactive agents.

## CCS CONCEPTS

• **Human-centered computing** → HCI theory, concepts and models.

## KEYWORDS

*Comfortability*, Human-Agent Interaction, Affective Computing, Multimodal Emotion Recognition


*ICMI '22, November 7–11, 2022, Bengaluru, India*

ACM ISBN 978-1-4503-9390-4/22/11…$15.00
https://doi.org/10.1145/3536221.3556631

**ACM Reference Format:**
Maria Elena Lechuga Redondo, Alessandra Sciutti, Francesco Rea, and Radoslaw Niewiadomski. 2022. *Comfortability* Recognition from Visual Non-verbal Cues. In *INTERNATIONAL CONFERENCE ON MULTIMODAL INTERACTION (ICMI '22), November 7–11, 2022, Bengaluru, India.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3536221.3556631

## 1 INTRODUCTION

Interactions entail a tangled mix of emotional, affective and internal states that emerge between the people who are communicating [11]. As a consequence, identifying others' internal states plays a very relevant role within social contexts [5, 22]. For this reason, any interactive agent would greatly benefit from being socially intelligent [42, 49]. Given that developing fully socially intelligent agents is a challenge beyond our reach, providing them with foundational skills could already have a positive impact on human-agent exchanges. Hence, this paper tackles one of these basic skills: *Comfortability* detection.

*Comfortability* was introduced in [45] as "*(disapproving of or approving of) the situation that arises as a result of an interaction, which influences one's own desire of maintaining or withdrawing from it*". The strong point about *Comfortability* is that it focuses on how a person feels respect to other agents' actions without deepening on the specific emotional or affective states that might arise in parallel. This way, a system capable of identifying someone's *Comfortability* would be able to understand whether it has acted appropriately and accordingly to its user's expectations and could assess whether it needs to adapt its behaviour. Albert Mehrabian established in 1967 the 7%–38%–55% rule, declaring that the *7%* of the communication is verbal, *38%* of the communication is vocal and *55%* of the communication is visual [38]. This statement justifies the importance of non-verbal communication highlighting at the same time how relevant is to be capable of understanding and recognizing others' nonverbal cues. Additionally, Maréchal et al. [35] wrote "A challenge in multi-modal emotion analysis is to efficiently explore emotion, not only on one but on highly expressive nature modalities." Therefore, this paper presents for the first time a model capable of classifying whether someone is *uncomfortable* or not, by paying attention to several non-verbal features. Given that the face is one of the most expressive modalities [5], different cues associated to it (e.g., *Action Units (AUs) and Gaze*), in addition to *Upper Body* and *Head Pose* cues, have been approached. Concisely, different Feature-based Machine Learning (ML) algorithms (i.e., *Naive Bayes (NB), Neural Networks (NN), Random Forest (RF)*

and *Support Vector Machines (SVM)*) have been fed with such features. Furthermore, the features under study were automatically extracted from spontaneous reactions recorded during several real Human-Robot Interaction (HRI) interviews.

This paper provides promising results together with ideas that might enhance subsequent *Comfortability* classifiers that at the same time, might help non-human agents to better understand their human partners.

## 2 LITERATURE REVIEW

### 2.1 Expressing and Perceiving Internal States through Non-verbal cues

One of the main channels to express and perceive emotional and affective states is the face. Almost all interactive agents have a face, and probably for this reason, humans are capable of identifying faces within the first few days after birth [48]. Barrett et al. [5] have deeply explored this area and explained that the concept "emotion" refers to a category of instances that vary from one another in their physical (e.g., facial and body movements) and mental (e.g., pleasantness, arousal, etc.) features. This way, they implied that an emotion (e.g., anger) won't own characteristics that are identical across situations, people and cultures. Conversely, Paul Ekman [18] defends that "there is a core facial configuration that can be used to diagnose a person's emotional state in the same way that a fingerprint can be used to uniquely recognize a person"[5]. He defines emotions as "a process, a particular kind of automatic appraisal influenced by our evolutionary and personal past, in which we sense that something important to our welfare is occurring, and a set of physiological changes and emotional behaviours begin to deal with the situation" [18]. In his book [18], he added that words are one way to deal with emotions, and that even though we use words when we are emotional, we cannot reduce emotions to words. Together with Friesen [20], both studied how people from an isolated tribe in New Guinea who had not interacted with anyone from outside their tribe, expressed and perceived each one of the called six-basic-emotions (i.e., *anger, surprise, fear, happiness, disgust* and *sadness*). To analyze their expressiveness, Ekman and Friesen defined several stories (associated to each one of the basic emotions) and asked the New Guineans to imagine themselves in such situations. Their faces were recorded and given to American collaborators, who had never traveled to New Guinea or been in contact with people from this tribe, to classify them into one of the *six-basic-emotions*. The results showed that the American collaborators were able to correctly classify all the videos except the ones associated with *fear* and *surprise*, which were interchangeably classified. To understand the New Guinean's perceptive abilities, Ekman prepared another experiment. This time, the New Guineans had to associate a story (which Ekman read to them) to a picture of a Caucasian face posing one of the *six-basic-emotions*. This experiment was performed with more than 300 people. The results showed that the subjects were very good in identifying *happiness, anger, disgust* and *sadness*. However, similarly to the Americans, they were unable to distinguish *fear* from *surprise*. They argued that this phenomenon might be due to not-well-formulated stories, but also to the fact that *fear* and *surprise* may be often intermingled in these people's lives, and thus not distinguished. Both studies provided evidence in favour of assuming that there are innate facial movements associated to standards and reproducible emotional states among situations and cultures.

Even though Barrett's and Ekman's positions offer evidence against each other, both provide rich information about how emotions are expressed and perceived by people. Barret stated that when we see someone performing a facial movement (e.g., smiling) and subsequently infer that that person is in an specific emotional state (e.g., happy) we are assuming that the smile reveals something about the person's internal state which cannot be accessed directly. This skill requires calculating a conditional probability of that person being in a particular internal state given the observable set of features (in this case facial features). This approach is not different from how machine learning systems operate to recognize emotions, even though as humans we do it without realizing it and constantly. Ekman and Friesen avoided the issue of associating specific facial configurations to specific internal states by creating/expanding the **Facial Action Coding System (FACS)** in 1978 [19], originally introduced by Hjortsjö in 1970 [25]. This system describes all the visual discernible facial movements, breaking down facial configurations into individual components of muscle movements, called **Action Units (AUs)**. Hjortsjö explored 23 AUs and afterwards, Ekman and Friesen expanded it to 64. To the date, there are 46 AUs which consider facial movements, 8 which consider head movements and 4 focused only on eye movements. Additionally, Baltrušaitis et al. developed Open Face [3], a software that automatically detects some AUs. Concretely, OpenFace is capable of detecting: *AU1, AU2* and *AU4* (which represent the muscle movements around the eyebrows), *AU5, AU6*, and *AU7* (which represent the muscle movements around the eyes) , *AU9* (which represents a nose wrinkle), *AU10, AU12, AU14, AU15, AU17, AU20, AU23, AU25* and *AU26* (which represent the muscle movements around the mouth) and *AU45* (which represents the action of blinking).

In spite the face has been deeply explored over the years, there are other channels that can reveal plenty of information about our internal states as well. One clear example is the body, which transmits a huge amount of information and has also been explored by Ekman [16] and other researchers who are trying to exploit its full-potential. For example, Hidalgo et al. [24] developed OpenPose, an automatic recognition software that automatically detects corporal poses distributed along the whole body, providing precise information about the face, hands, and feet. Another good example is the voice. Human auditory information (i.e., *pitch, timbre, loudness*, and *vocal tone*) has been proven to express emotions during speech generation [12]. To this point, all the introduced emotional channels are perceivable by the human senses up to a certain extent. Cacioppo et al [7] affirmed that human's affective response is a psycho-physiological process triggered by stimuli, which is often manifested through observable behaviour channels. Although not all physiological signals can be efficiently perceived through our senses, their changes can be measured with technological devices. This means that, even though we might not be able to use this information on daily interactions, we might want to evaluate someone's internal states by considering also these features. For example, Lobbestael et al. [30] conducted an study focused on anger, where they exposed sixty-four participants to one specific stimulus (either a movie, a stressful interview, punishment or harassment).

To measure their anger, they considered self-reports and a list of physiological signals: *blood pressure, heart rate, skin conductance level,* and *skin conductance response*. They found that all the stimuli produced similar self-reports, but that their cardiovascular effects, and electrodermal activity increased more during the harassment and stressful interview. This might suggest again that people might "control" what they verbally say, or how to voluntarily behave, however they cannot control how their hearth will beat or their body will sweat. Hence, physiological signals might bring extremely useful information to establish ground-truths concerning internal states.

## 2.2 What should affective computing compute?

Most research in affective computing tackles the well-known "six-basic-emotions" (i.e., *happiness, sadness, anger, fear, disgust* and *surprise*). In fact, the most popular databases are based on them (e.g., the JAFFE[33, 34], KDEF[32], BU-3DFE [54], CK+ [31], MMI [41], SPEF [13], EMOTIONNET [21], AffectNet [40], and RAF-AU [53] databases) and plenty of researchers have developed algorithms capable of classifying them obtaining promising results [43, 51]. Nonetheless, "non-basic-emotions" (e.g., *engagement, boredom, confusion, frustrations* and so on) were found to be five times more frequent in real-live situations [14]. It makes sense that six basic emotions might be insufficient to cover all the complex feelings and feedback felt and expressed during social situations. At the same time, almost all the expressions contained in these databases are acted, which means that they might not be a real reflection of the expressions that arise in real life. To date, there are some popular databases which include spontaneous not-acted data (e.g., the DISFA [37] and BP4D[57] databases), however they still present a lack of more complex internal states.

In general, it can be noticed that there is a need of databases that contain not only spontaneous reactions, but internal states that emerge during daily situations. This way, future affective computing would address better human-machine interactions. As a contribution to this idea, this paper provides a start by developing a *Comfortability* recognition system based on genuine human behaviour.

## 2.3 Machine Learning Algorithms

To automatically recognize any aspect of communication (e.g., an emotion/internal state), a Machine Learning (ML) algorithm is usually designed and trained. On the basis of the input information, ML algorithms can be divided into two branches: *Deep Learning* (when the algorithm is capable of processing information without any previous computation) and *Feature-based Learning* (when the algorithm receives a set of pre-processed features). At the same time, ML algorithms can be characterized as *supervised* or *unsupervised* learning depending on its classification strategy. Supervised ML algorithms require labeled data and the latter do not, as they autonomously identify clustering principles.

Recent studies focused on affective computing have considered different approaches and modalities. For example, Rajan et al. [43] created a model based on *Convolutional Neural Network* (CNN) and *Long-Short Term Memory* (LSTM) that would take into account dynamic temporal information for facial expression recognition. They tested the model with well-known databases (CK+, MMI, and SPEW), obtaining an accuracy of 99%, 80% and 56% respectively. For some reason, the classes *anger, fear* and *sadness* were worse classified than the others. In the same fashion, Bartlett et al. [6] developed a conceptor based low/high engagement classifier based on *Recursive Neural Networks (RNN)*. To feed the classifier, they extracted skeletal and facial landmarks using the OpenPose [8] software from the videos contained in the PInSoRo [29] data-set (children performing tasks) taking into account the movement. They obtained a recognition accuracy of 60% for the clips annotated as High Engagement and a recognition accuracy of 75% for the clips annotated as Low Engagement. Castellano et al. [9] also studied the role of movement when inferring emotions. To do so, they used videos collected during the third summer school of the Human-Machine Interaction Network on Emotion (HUMAINE) EU-IST project held in Genova in 2006. In particular, they used 240 dynamic gestures of 8 different emotions (*anger, despair, interest, pleasure, sadness, irritability, joy* and *pride*) acted by 10 different actors. They represented each movement by computing its *Quantity of Motion (QoM), Contraction Index, Velocity, Acceleration* and *hand's fluidity barycenter*. Then, they applied a Dynamic Time Wrapping (DTW) [28] algorithm to measure similarities between movements. After comparing the five corporal features, they learnt that QoM was the one with lower classification error when distinguishing between *anger, joy, pleasure* and *sadness*. The remaining emotions were found to be unsuccessfully classified by any of the proposed features. One last example relevant to this paper is Matsufuji et al. [36] who developed a model to detect awkward situations. They considered voice intonation (i.e., maximum pitch and speech length) and corporal information extracted with the Kinect sensor (i.e., head pitch, yaw, neck, shoulder and elbow velocity vectors; and head x and z axes) of 5 subjects. They used these features with the Weka [15] software and several ML algorithms. They obtained a recognition accuracy of 83% for *Bayesian Networks*, 72 % *Random Forest*, 72% *Support Vector Machines*, and 70 % for *Naive Bayes*.

As literature shows, there are plenty modalities and algorithms that can be considered. We agree with literature that the more modalities present (e.g., physiological, auditory, visual, etc.), the more likely the model's performance will improve. Nonetheless, as this paper presents an initial approach to build a *Comfortability* classifier, it was decided to tackle one aspect at a time. On the one hand, given the long-term aim of this project is to build a system capable of working in ecological scenarios (i.e., where no external devices are placed on the subject), physiological data were not considered. On the other hand, we noticed (analyzing the recordings) that other modalities (i.e., body movements, audio (e.g., pitch and tone), context and verbal content (i.e., the use of verbs)) seemed to be relevant to represent *Comfortability*. However, we observed that some of them might be quite challenging to interpret as people might desire to hide their *Uncomfortability* with verbal statements and/or feel different under similar circumstances. Hence, we decided to focus on the facial and upper body information, leaving for further studies the other features. Regarding the ML algorithm, we decided to explore several Feature-based Learning ones passing them positions, velocities and AUs (as they do not rely on faces' contours, colors, gadgets and hairstyles) as a first attempt. More

complex features (e.g., *QoM*) as well as DL approaches are tentative candidates for future studies.

## 3 METHODS

To capture spontaneous and legit reactions within a Human-Machine Interaction scope, the iCub robot [39] interviewed several researchers for a real and novel column of our institutional online magazine [1] [44]. During the interviews, the participants were exposed to an stressful interaction where the robot *complimented* them at the beginning and *interrupted, ignored* and *misunderstood* them at the end. Even though plenty of data (*auditory*, *visual* and *physiological*) were collected, only the *visual information* is explored in this paper.

A total of **29 videos** (one per interview) of 17 : 54(±5 : 17 SD) minutes on average were recorded. From those videos, only 26 were used for this study, because 3 participants were interviewed from a different perspective (instead of a total frontal view, they were recorded slightly turned, like a classical TV interview). Our data-set is peculiar, not only because the reactions are provoked by a non-human agent, but because our participants are from very different cultures and ethnicities; which to date is rare to find [5]. To analyze the visual information, the audio was excluded from the videos with the intention of not allowing the annotator discover the context and hence, be biased. Afterwards, the videos were trimmed into smaller segments and subsequently labelled.

### 3.1 Preparing our Data-set - Trimming and Labelling

Reis et al. [46] wrote that "the most fundamental property of a coding scheme for observing social interactions is the technique adopted for sampling behaviour, known as *unitizing*". Unitizing means dividing an observable sample into discrete smaller samples. According to cognitive sciences [55], unitizing is an automatic component of the human perceptual processing of the ongoing situation. That is to say, we as humans make sense of reality by breaking it into smaller units. Ceccaldi et al. [10] added that artificial agents should master unitizing skills to reach a comprehensive understanding of the interaction itself. With that goal in mind, they explored the drawbacks and benefits of the two main unitizing techniques (*Interval* and *Continuous* coding). On the one hand, *Interval Coding* consists of identifying a fixed-length time interval in which the sample will be segmented into. It is expected that the raters should be able to find occurrences of the targeted behaviour in those pieces. Established research [2] proved that thin slices (i.e., from *2 seconds to 5 minutes*) is a well-known approach for personality, affect and interpersonal relationship samples. Even though this technique might cut actions in between and thus, relevant information can be lost, it is fast, easy to automatize, objective and there is no need of a prior knowledge of the context. On the other hand, *Continuous Coding* stands for identifying specific behaviours that are likely to last different amounts of time, where each segment will have its own size. While this technique comprehends exactly the desired information, it is much more time consuming and often requires trained annotators. Moreover, it is likely that establishing a continuous segmentation will require a coding scheme itself (there are some predefined like ACT4Teams [27]). Regarding our samples, we initially thought of using a *Continuous Coding*, given that the observed behaviors seemed to not be constant and hence, vary in time. We started by looking at each clip trying to isolate each facial configuration (which by itself could represent a particular *Comfortability* level) from others. Nevertheless, after several attempts performing a customized isolation of these movements, we realized that this was not effective. Identifying the beginning and ending point of a unitary facial movement was not trivial and required to consider both facial movements and other complex features (e.g., facial skin color). Furthermore, the specific moment a facial movement started and ended could be perceived differently by different people and even by the same person at different times. In fact, Afzal et al. reached the same conclusions [1]. After annotating a data-set which contained spontaneous, unpredictable and natural reactions, they concluded *"Even for a human expert, it is difficult to define what constitutes an emotion. Segmenting the original videos into emotionally salient clips was the most labour-intensive and time-consuming process. Demarcating the beginning and end of emotional expressions is incredibly hard as they often overlap, co-occur or blend subtly into a background expression"*. This aspect made us believe that segmenting our data-set following a continuous segmentation would require an experimental set-up on its own; which is not the focus of our research. Therefore, we decided to go with an *Interval Coding* even though some expressions might be cut in between. To avoid cutting movements related to different events, a two layer segmentation process was performed. The first step was to segment each video into 24 segments; one per each relevant interview part associated to a *Comfortability* level (also in line with the self-report's structure). The second step was to segment each one of those segments into smaller pieces. To decide the length of each piece, the time a macro-expression (from 1/2 to 4 seconds) and micro-expression (from 1/2 to 1 second) tend to last [17] were taken into account. Therefore, each one of those 24 segments (counting all the participants, a total of 696 segments) was segmented again into *3-seconds* segments. The final amount of prepared segments was of 10.468 units. The segments which were shorter than *3-seconds*, as a result of the trimming, were discarded leaving a final amount of 8.467 units.

Afterwards, each one of the three-second sample was labelled following a 7-point Likert scale from *1* (being *Extremely Uncomfortable*) to *7* (being *extremely Comfortable*). Annotate a sample, judge each participant's response is a very challenging and demanding task. The mood and fatigue of the annotator, as well as the previously annotated sample can bias the evaluation criteria inducing to error and subjectivity [47]. Also, it is known that facial movements can be consciously shaped. For example, the *"Duchenne Smile"* might be interpreted at first sight as a "happiness" indicator. However, it was found that it can be intentionally produced to signal submission or affiliation [23]. In addition to this finding, Hoque et al. [26] performed an experiment to study friendly vs. polite smiles. The experiment consisted of people interested in banking services meeting with a professional banker face-to-face. They discovered that amused smiles present themselves longer and more symmetrical than those enacted out of politeness. In addition, as it can happen during *unitizing*, an annotator can become an expert by performing several rounds of annotations learning after each repetition a particular aspect of the emotional response or

[1] https://opentalk.iit.it/i-got-interviewed-by-a-robot/

cue under study. Hence, trying to minimize the weak-spots, we became experts by running three annotation rounds and annotated the data-set. During the annotation, the three-second videos were presented one at a time in the screen with a 1920x1080 resolution using the **MUltiple VIdeos LABelling (MuViLab) annotation tool** [2] software. Once a video appeared, it was played in loop, allowing the annotator to introduce a *Comfortability* level from *1* to *7* by pressing a number in the keyboard, until the annotator decided to pass to the next one. The clips were presented in a random order, which prevented the annotator familiarizing with one specific person and understanding the specific context of the expression under analysis.

## 3.2 Non-Verbal Features

A set of **118 features** were extracted from the three-second videos and considered for the *Comfortability* classifier. In particular, four different algorithms: *Naive Bayes (NB), Neural Networks (NN), Random Forest (RF)* and *Support Vector Machines (SVM)* were trained and tested with the following features:

*3.2.1 AUs.* The person's Action units (AUs) were extracted using the OpenFace software [3]. Specifically, the Action Unit *AU01, AU02, AU04, AU05, AU06, AU07, AU09, AU10, AU12, AU14, AU15, AU17, AU20, AU23, AU25, AU26* and *AU45*; which are the ones the software recognizes. The mean and standard deviation of each intensity of the sixteen AUs were included for each three-second video clip. Thus, a total of **34 features associated to the person's Action Units** were included.

*3.2.2 BodyPose.* The person's corporal information was extracted using the OpenPose software [8, 50]. Specifically, information about the person's upper body positions (i.e., *x* and *y* of 3 key-points per arm, 1 in between the shoulders, and 5 in the head). The mean and standard deviation of each one of these key-points coordinates was computed per each three-second clip. Thus, a total of **48 features associated to the person's upper body** were included.

*3.2.3 Gaze.* The person's gaze was extracted using the OpenFace software [52]. Specifically, the eye gaze direction vector in world coordinates for each eye (i.e., the *x, y* and *z* coordinates for the *left eye* and the *x, y* and *z* coordinates for the *right eye*), the eye gaze angle direction averaged for both eyes. The mean and standard deviation of all these features were considered per each three-second video clip. Thus, a total of **16 features associated to the person's gaze** were included.

*3.2.4 HeadPose.* The person's head position and rotation extracted using the OpenFace software [4, 56]. Specifically, the location of the head with respect to the camera in millimeters (i.e., the *x, y* and *z* coordinates; where a positive Z is being further away from the camera) and the rotation of the head in world coordinates with the camera being the origin (i.e., the *x, y* and *z* coordinates representing the pitch, yaw and roll respectively). The mean, standard deviation, velocity and acceleration of the head location and rotation were considered when creating the classifier during each three-second video clip. Thus, a total of **20 features associated to the person's head location and movement** were included.

[2] *github.com/ale152/muvilab*

# 4 RESULTS

In order to build a *Comfortability* model capable of classifying whether someone is *uncomfortable*, several ML algorithms were developed, where each algorithm's variable was tuned to its optimal performance for each feature received as input. More details are provided in the subsequent sections. Also, even though some of the features associated to a specific three-second clip took into account temporal dynamics, the algorithms did not consider a sequence between clips. Thus, the data did not follow the interview sequence. In addition, the data were divided into 70% per training (with a 30% used for cross-validation) and 30% per testing (see Table 1). The clips were not discriminated among subjects, which means that a clip reserved for testing was not seen during the whole training procedure, but the subject was. The algorithm with the best accuracy was also tested with a leave-one subject out approach. Figure 1 shows the percentage of clips annotated with each one of the 7 *Comfortability* levels. It can be seen that the *Comfortability* extremes are poorly represented, being those barely 4% of the data-set. Nevertheless, the data were appreciably balanced while splitting the samples into *being Not-Uncomfortable* (i.e., being comfortable or neutral) 51% and *being Uncomfortable* 49%. Table 1 includes the specific number of clips used for training and testing for each subsequent *Comfortability* label.

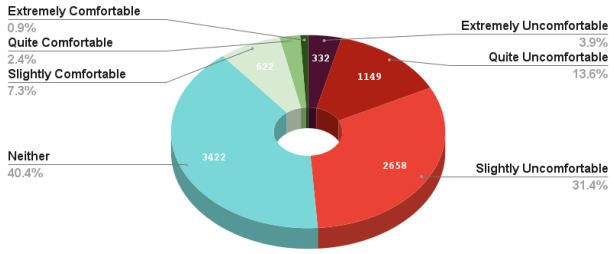


**Figure 1: Percentages of video clips annotated for each *Comfortability* level**

| *Comfortability* Label | # Clips Training | # Clips Testing |
|---|---|---|
| Not-Uncomfortable | 3020 | 1309 |
| Uncomfortable | 2907 | 1230 |

**Table 1: Number of video clips used to train and evaluate the interviewees' *Comfortability***

## 4.1 Naive Bayes

Table 2 shows the *accuracy*, *precision* and *recall* for different combinations of features used to train and evaluate the **Naive Bayes classifier**. From these results, it can be seen that the AUs together the Gaze are the ones that the algorithm learns better during training obtaining a 69% accuracy. When evaluating the model with unseen data, AUs together BodyPose are the features that work better, obtaining a *Comfortability* accuracy of almost 65%.

| Input | Training-set Accuracy | Testing-set Accuracy | Precision | Recall |
|---|---|---|---|---|
| AUs | .603 | .591 | .595 | .587 |
| BodyPose | .582 | .580 | .607 | .587 |
| Gaze | .565 | .570 | .585 | .562 |
| HeadPose | .510 | .515 | .757 | .500 |
| AUs + BodyPose | .649 | **.649** | .649 | .648 |
| AUs + Gaze | **.691** | .594 | .600 | .589 |
| AUs + HeadPose | .510 | .515 | **.757** | .500 |
| BodyPose + Gaze | 607 | .626 | .631 | .628 |
| BodyPose + HeadPose | .516 | .522 | .634 | .507 |
| Gaze + HeadPose | .510 | .515 | .757 | .500 |
| AUs + BodyPose + Gaze | .644 | .647 | .649 | .645 |
| AUs + BodyPose + HeadPose | .516 | .522 | .634 | .507 |
| AUs + Gaze + HeadPose | .510 | .515 | **.757** | .500 |
| BodyPose + Gaze + HeadPose | .516 | .522 | .634 | .507 |
| AUs + BodyPose + Gaze + HeadPose | .516 | .522 | .634 | .507 |

**Table 2: Naive Bayes *Comfortability* classification considering the features extracted from ecological three-second clips**

## 4.2 Neural Networks

Table 3 shows the *accuracy*, *precision* and *recall* for different combinations of features used to train and evaluate the ***MLPClassifier*** **Neural Networks classifier** of *sklearn*. To obtain the best accuracy, the classifier was tuned for each specific input, varying its activation function (*identity, logistic, tanh* or *relu*), solver (*lbfgs, sgd* and *adam*) and hidden layers' size (from *1* to *35* layers) until obtaining its maximum accuracy. As a result, the model trained with AUs and HeadPose features obtained the highest training-set performance with more that 78% accuracy. On top of that, a combination of AUs , BodyPose and HeadPose features, and a *logistic*, *adam* and *30-hidden-layers* configuration led to the highest performance with unseen data, obtaining a 72% *Comfortability* recognition accuracy.

## 4.3 Random Forest

Table 4 shows the *accuracy*, *precision* and *recall* for different combinations of features used to train and evaluate the **Random Forest classifier**. From these results, it can be seen that all the features and combination of features performed perfectly with the training-set. Additionally, it was found that a combination of all the features is the best bet for this algorithm with unknown data. Merging the AUs, BodyPose, Gaze and HeapPose features enhanced the model to a 75% *Comfortability* recognition *accuracy*, *precision* and *recall*.

## 4.4 Support Vector Machines

Table 5 shows the *accuracy*, *precision* and *recall* for different combinations of features used to train and evaluate the **Support Vector Machines classifier**. For each input feature/s, all possible combinations of kernels (*linear, polynomial, rbf* and *sigmoid*), C (from *.001* to *100*) and gamma (from *.001* to *100*) values were run, choosing the one with the best accuracy. As a result, most of the inputs perform ideally with data already seen. On the other hand, AUs together with Gaze are the features that better represent someone's *Comfortability* level in the testing-set, reaching a 71% recognition accuracy. This model was trained with a *Radial Basis Function (rbf)* as *kernel* with *gamma* = .1 and *C* = 20.1 values.

## 4.5 Best Algorithms and Features

The combination of features which led to the best classification *accuracy* for each one of the tested ML algorithms is shown in Table 6. Looking at the training set, the RF and SVM algorithms are the ones that perform better, reaching a perfect recognition response. Considering the test set NB, NN and SVM do not perform very differently, while RF remains the one with the highest results. To explore deeply its performance, Figure 2 reports the classification performed on the **training data-set** and Figure 3 reports the classification performed on the **testing data-set**. As it can be noticed, the algorithm recognizes "Not-Uncomfortable" levels slightly better (77% of the time) than it recognizes "Uncomfortable" levels (73% of the time).

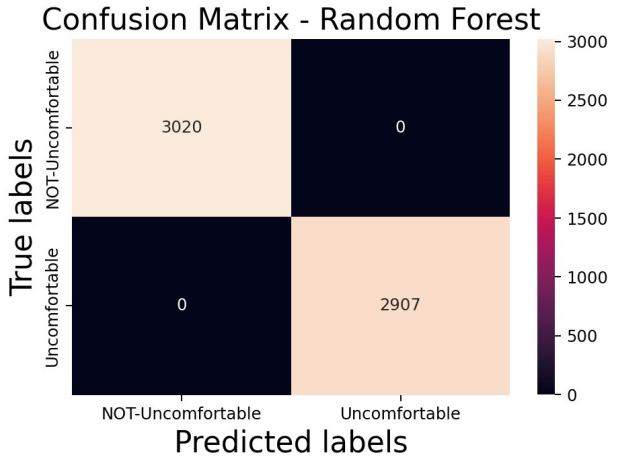


**Figure 2: Training-set *Comfortability* classification accuracy**

The algorithm with the best *accuracy* was also tested with a leave-one-subject-out procedure. Therefore, the model was trained 26 different times, each one leaving one subject out of the training set to test the final system accuracy with it. This way, the system is

| Input | Training-set Accuracy | Testing-set Accuracy | Precision | Recall |
|---|---|---|---|---|
| AUs | .747 | .682 | .683 | .680 |
| BodyPose | .712 | .684 | .684 | .683 |
| Gaze | .694 | .673 | .673 | .673 |
| HeadPose | .641 | .648 | .650 | .645 |
| AUs + BodyPose | .702 | .687 | .688 | .686 |
| AUs + Gaze | .745 | .713 | .713 | .713 |
| AUs + HeadPose | **.782** | .703 | .703 | .703 |
| BodyPose + Gaze | .530 | .523 | .594 | .535 |
| BodyPose + HeadPose | .705 | .679 | .679 | .678 |
| Gaze + HeadPose | .641 | .648 | .651 | .645 |
| AUs + BodyPose + Gaze | .722 | .697 | .697 | .697 |
| AUs + BodyPose + HeadPose | .740 | **.720** | .720 | .720 |
| AUs + Gaze + HeadPose | .760 | **.720** | .721 | .721 |
| BodyPose + Gaze + HeadPose | .716 | .685 | .685 | .683 |
| AUs + BodyPose + Gaze + HeadPose | .745 | .706 | .706 | .704 |

**Table 3: Neural Networks *Comfortability* classification considering the features extracted from ecological three-second clips**

| Input | Training-set Accuracy | Testing-set Accuracy | Precision | Recall |
|---|---|---|---|---|
| AUs | 1 | .700 | .700 | .700 |
| BodyPose | 1 | .739 | .739 | .739 |
| Gaze | 1 | .667 | .667 | .666 |
| HeadPose | 1 | .724 | .724 | .724 |
| AUs + BodyPose | 1 | .749 | .749 | .748 |
| AUs + Gaze | 1 | .715 | .715 | .716 |
| AUs + HeadPose | 1 | .737 | .737 | .737 |
| BodyPose + Gaze | 1 | .744 | .744 | .743 |
| BodyPose + HeadPose | 1 | .738 | .738 | .738 |
| Gaze + HeadPose | 1 | .739 | .738 | .738 |
| AUs + BodyPose + Gaze | 1 | .745 | .745 | .745 |
| AUs + BodyPose + HeadPose | 1 | .747 | .747 | .746 |
| AUs + Gaze + HeadPose | 1 | .740 | .740 | .740 |
| BodyPose + Gaze + HeadPose | 1 | .747 | .747 | .746 |
| AUs + BodyPose + Gaze + HeadPose | **1** | **.752** | .751 | .751 |

**Table 4: Random Forest *Comfortability* classification considering the features extracted from ecological three-second clips**

tested on subjects it was not trained on. As a result, the **Random Forest** classifier was trained with a combination of the subject's *AUs*, *BodyPose*, *Gaze* and *HeadPose* features obtaining a classification accuracy average of 56.6%(±14.2% SD). Paying attention to the individual subjects, it is observed that not everyone was classified with the same *accuracy*. While some obtained very poor results (from 27% to 47%) others achieved quite nice performances (from 53% to 81%). The difference between the classification accuracy of the *testing-set* procedure (75%) and the *leaving-one-subject-out* procedure (57%) might be due to the subjects' sample size. That is to say, being highly likely that people express *Comfortability* in their own manner, a system not familiar with a particular person might have it extremely complicated to understand what being *Uncomfortable* or *Not-uncomfortable* means; i.e., how people behave when being *Uncomfortable* or *Not-uncomfortable*. Both systems were only trained with 26 subjects (counting the one/s used for testing). Instead, if the models were to be fed with many more subjects, the more likely they would encounter people that express common *Comfortability* patterns and thus the better the model would classify data from unknown subjects. For this reason, it has been particularly challenging for the *leaving-one-subject-out* model to generalize and predict how an unknown person would express their own *Comfortability*. In spite of that, the model is capable of classifying data it has never been exposed to better than chance.

## 5 DISCUSSION AND FUTURE WORK

This paper has presented several ML models capable of recognizing *Comfortability* by taking into account different non-verbal cues that arose during a real interaction between a person a humanoid robot. Specifically, the features under study comprehended information about the participants' facial and upper body movements (i.e., Action Units (AUs), Head Position, Gaze and Upper-body Position). The best algorithm was trained with a combination of all the proposed features obtaining a 75% accuracy. At the same time,

| Input | Training-set Accuracy | Testing-set Accuracy | Precision | Recall |
|---|---|---|---|---|
| AUs | .717 | .686 | .687 | .684 |
| BodyPose | 1 | .529 | .641 | .514 |
| Gaze | .614 | .620 | .622 | .616 |
| HeadPose | .999 | .586 | .619 | .577 |
| AUs + BodyPose | 1 | .528 | .638 | .514 |
| AUs + Gaze | .748 | **.709** | .709 | .708 |
| AUs + HeadPose | **1** | .584 | .618 | .575 |
| BodyPose + Gaze | 1 | .529 | .641 | .515 |
| BodyPose + HeadPose | 1 | .517 | .604 | .502 |
| Gaze + HeadPose | .998 | .613 | .624 | .608 |
| AUs + BodyPose + Gaze | 1 | .528 | .638 | .514 |
| AUs + BodyPose + HeadPose | 1 | .516 | .576 | .501 |
| AUs + Gaze + HeadPose | .998 | .623 | .637 | .617 |
| BodyPose + Gaze + HeadPose | 1 | .519 | .591 | .504 |
| AUs + BodyPose + Gaze + HeadPose | 1 | .518 | .586 | .503 |

**Table 5: SVM *Comfortability* classification considering the features extracted from ecological three-second clips**

| Algorithm | Input | Accuracy Train/Test |
|---|---|---|
| Naive Bayes | AUs + Gaze | 69% / 60% |
| Neural Networks | AUs + HeapPose | 78% / 70% |
| **Random Forest** | **AUs + BodyPose + Gaze + HeadPose** | **100% / 75%** |
| Support Vector Machines | AUs + HeadPose | 100% / 59% |

**Table 6: Best algorithms performance considering the whole data-set and 2 *Comfortability* labels: being *Uncomfortable* vs being *Not-uncomfortable***

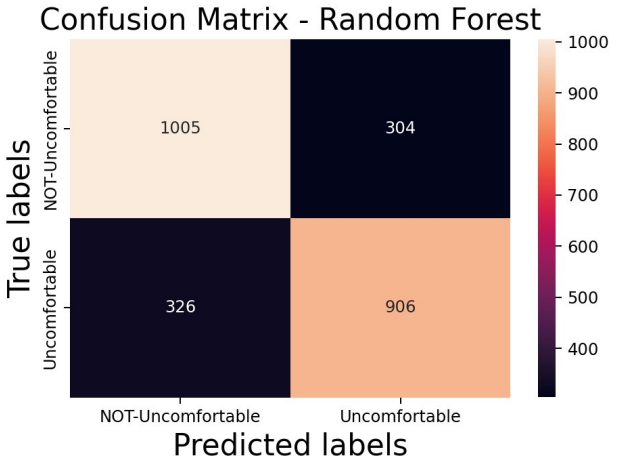


**Figure 3: Testing-set *Comfortability* classification accuracy**

the same architecture was evaluated leaving one subject out during training to test with it. This decreased the accuracy obtained before, but still maintained a recognition accuracy better than chance (i.e., 58%). This means that the model is capable of recognizing whether someone is uncomfortable or not, not only of people that has already interacted with, but of total unknown faces that has not even seen. It has been proven that, even though it is not the case for all the four ML algorithms explored, the more features are combined the more accurate predictions would be produced by the model. Bearing in mind this thought, it is likely that the classifier presented in this paper could be enhanced if more modalities would be taken into account. As mentioned before, synchronized *audio*, *video* and *physiological signals* have been recorded, together with the *context* (the type of question being asked at the precise time). Future steps could focus on exploring deeply and individually each one of these features, and then merging them together, to discover how to best combine them to build an effective *Comfortability* Artificial Intelligence. Additionally, the features used to feed the system can be polished and selected. At the moment, averages and standard deviations have been used to represent the temporal dynamics and static positions of the aforementioned features. However, some of these features might be poorly or redundantly represented. For these reasons, more complex features (like the *contraction index* of the expression, *quantity of motion* (QoM) of the subject and so on) and dimensionality reduction techniques like PCA should be computed and applied to improve the model. In the same fashion, more complex models (possible Deep-Learning based) could be considered.

Another very important aspect that might improve considerably the model's performance regarding unknown faces, is to expand the data-set by collecting more videos from a bigger sample of subjects. Given that people express internal states differently, a much more varied data-set could increase the chance of recognizing *Comfortability* levels expressed in several unexpected ways.

Overall, this study has presented an accurate *Comfortability* recognition system, and highlighted relevant factors that might improve the *Comfortability* classifier considerably.

## ACKNOWLEDGMENTS

Alessandra Sciutti is supported by a Starting Grant from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program. G.A. No. 804388, wHiSPER.

## REFERENCES

[1] Shazia Afzal and Peter Robinson. 2011. Natural affect data: Collection and annotation. In *New perspectives on affect and learning technologies*. Springer, 55–70.
[2] Nalini Ambady and Robert Rosenthal. 1992. Thin slices of expressive behavior as predictors of interpersonal consequences: A meta-analysis. *Psychological bulletin* 111, 2 (1992), 256.
[3] Tadas Baltrušaitis, Marwa Mahmoud, and Peter Robinson. 2015. Cross-dataset learning and person-specific normalisation for automatic action unit detection. In *2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, Vol. 6. IEEE, 1–6.
[4] Tadas Baltrusaitis, Peter Robinson, and Louis-Philippe Morency. 2013. Constrained local neural fields for robust facial landmark detection in the wild. In *Proceedings of the IEEE international conference on computer vision workshops*. 354–361.
[5] Lisa Feldman Barrett, Ralph Adolphs, Stacy Marsella, Aleix M Martinez, and Seth D Pollak. 2019. Emotional expressions reconsidered: Challenges to inferring emotion from human facial movements. *Psychological science in the public interest* 20, 1 (2019), 1–68.
[6] Madeleine Bartlett, Daniel Hernandez Garcia, Serge Thill, and Tony Belpaeme. 2019. Recognizing human internal states: a conceptor-based approach. *arXiv preprint arXiv:1909.04747* (2019).
[7] John T Cacioppo, Louis G Tassinary, and Gary Berntson. 2007. *Handbook of psychophysiology*. Cambridge university press.
[8] Z. Cao, G. Hidalgo Martinez, T. Simon, S. Wei, and Y. A. Sheikh. 2019. OpenPose: Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2019).
[9] Ginevra Castellano, Santiago D Villalba, and Antonio Camurri. 2007. Recognising human emotions from body movement and gesture dynamics. In *International Conference on Affective Computing and Intelligent Interaction*. Springer, 71–82.
[10] Eleonora Ceccaldi, Nale Lehmann-Willenbrock, Erica Volta, Mohamed Chetouani, Gualtiero Volpe, and Giovanna Varni. 2019. How unitizing affects annotation of cohesion. In *2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE, 1–7.
[11] C. Clare. 2012. *Communicate : how to say what needs to be said when it needs to be said in the way it needs to be said.* National Library of Australia Cataloguing-in-Publication entry. 187 pages.
[12] Poorna Banerjee Dasgupta. 2017. Detection and analysis of human emotions through voice and speech pattern processing. *arXiv preprint arXiv:1710.10198* (2017).
[13] Abhinav Dhall, Roland Goecke, Simon Lucey, and Tom Gedeon. 2011. Static facial expression analysis in tough conditions: Data, evaluation protocol and benchmark. In *2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops)*. IEEE, 2106–2112.
[14] Sidney D'Mello and Rafael A Calvo. 2013. Beyond the basic emotions: what should affective computing compute? In *CHI'13 extended abstracts on human factors in computing systems*. 2287–2294.
[15] Frank Eibe, Mark A Hall, and Ian H Witten. 2016. The WEKA workbench. Online appendix for data mining: practical machine learning tools and techniques. In *Morgan Kaufmann*. Elsevier Amsterdam, The Netherlands.
[16] Paul Ekman. 1965. Differential communication of affect by head and body cues. *Journal of personality and social psychology* 2, 5 (1965), 726.
[17] Paul Ekman. 2003. Darwin, deception, and facial expression. *Annals of the new York Academy of sciences* 1000, 1 (2003), 205–221.
[18] P. Ekman. 2004. Emotions revealed. *Bmj* 328, Suppl S5 (2004).
[19] Paul Ekman and Wallace V Friesen. 1978. Facial action coding system. *Environmental Psychology & Nonverbal Behavior* (1978).
[20] Paul Ekman, E Richard Sorenson, and Wallace V Friesen. 1969. Pan-cultural elements in facial displays of emotion. *Science* 164, 3875 (1969), 86–88.
[21] C Fabian Benitez-Quiroz, Ramprakash Srinivasan, and Aleix M Martinez. 2016. Emotionet: An accurate, real-time algorithm for the automatic annotation of a million facial expressions in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 5562–5570.
[22] D. Golleman. 2006. *Social Intelligence: The Revolutionary New Science of Human Relationships*. Editorial Kairos. 544 pages.
[23] Sarah D Gunnery and Judith A Hall. 2014. The Duchenne smile and persuasion. *Journal of Nonverbal Behavior* 38, 2 (2014), 181–194.
[24] Gines Hidalgo, Yaadhav Raaj, Haroon Idrees, Donglai Xiang, Hanbyul Joo, Tomas Simon, and Yaser Sheikh. 2019. Single-network whole-body pose estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 6982–6991.
[25] Carl-Herman Hjortsjö. 1969. *Man's face and mimic language*. Studentlitteratur.
[26] Mohammed Hoque, Louis-Philippe Morency, and Rosalind W Picard. 2011. Are you friendly or just polite?–analysis of smiles in spontaneous face-to-face interactions. In *International Conference on Affective Computing and Intelligent Interaction*. Springer, 135–144.
[27] Simone Kauffeld, Nale Lehmann-Willenbrock, and Annika L Meinecke. 2018. 21 The Advanced Interaction Analysis for Teams (act4teams) Coding Scheme. (2018).
[28] Eamonn Keogh and Chotirat Ann Ratanamahatana. 2005. Exact indexing of dynamic time warping. *Knowledge and information systems* 7, 3 (2005), 358–386.
[29] Séverin Lemaignan, Charlotte ER Edmunds, Emmanuel Senft, and Tony Belpaeme. 2018. The PInSoRo dataset: Supporting the data-driven study of child-child and child-robot social dynamics. *PloS one* 13, 10 (2018), e0205999.
[30] Jill Lobbestael, Arnoud Arntz, and Reinout W Wiers. 2008. How to push someone's buttons: A comparison of four anger-induction methods. *Cognition & Emotion* 22, 2 (2008), 353–373.
[31] Patrick Lucey, Jeffrey F Cohn, Takeo Kanade, Jason Saragih, Zara Ambadar, and Iain Matthews. 2010. The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression. In *2010 ieee computer society conference on computer vision and pattern recognition-workshops*. IEEE, 94–101.
[32] Daniel Lundqvist, Anders Flykt, and Arne Öhman. 1998. Karolinska directed emotional faces. *Cognition and Emotion* (1998).
[33] Michael Lyons, Shigeru Akamatsu, Miyuki Kamachi, and Jiro Gyoba. 1998. Coding facial expressions with gabor wavelets. In *Proceedings Third IEEE international conference on automatic face and gesture recognition*. IEEE, 200–205.
[34] Michael J Lyons. 2021. " Excavating AI" Re-excavated: Debunking a Fallacious Account of the JAFFE Dataset. *arXiv preprint arXiv:2107.13998* (2021).
[35] Catherine Marechal, Dariusz Mikolajewski, Krzysztof Tyburek, Piotr Prokopowicz, Lamine Bougueroua, Corinne Ancourt, and Katarzyna Wegrzyn-Wolska. 2019. Survey on AI-Based Multimodal Methods for Emotion Detection. *High-performance modelling and simulation for big data applications* 11400 (2019), 307–324.
[36] Akihiro Matsufuji, Tatsuya Shiozawa, Wei Fen Hsieh, Eri Sato-Shimokawara, Toru Yamaguchi, and Lieu-Hen Chen. 2017. The analysis of nonverbal behavior for detecting awkward situation in communication. In *2017 Conference on Technologies and Applications of Artificial Intelligence (TAAI)*. IEEE, 118–123.
[37] S Mohammad Mavadati, Mohammad H Mahoor, Kevin Bartlett, Philip Trinh, and Jeffrey F Cohn. 2013. Disfa: A spontaneous facial action intensity database. *IEEE Transactions on Affective Computing* 4, 2 (2013), 151–160.
[38] Albert Mehrabian and Susan R Ferris. 1967. Inference of attitudes from nonverbal communication in two channels. *Journal of consulting psychology* 31, 3 (1967), 248.
[39] G. Metta, G. Sandini, D. Vernon, L. Natale, and F. Nori. 2008. The iCub humanoid robot: an open platform for research in embodied cognition. In *Proceedings of the 8th workshop on performance metrics for intelligent systems*. 50–56.
[40] Ali Mollahosseini, Behzad Hasani, and Mohammad H Mahoor. 2017. Affectnet: A database for facial expression, valence, and arousal computing in the wild. *IEEE Transactions on Affective Computing* 10, 1 (2017), 18–31.
[41] Maja Pantic, Michel Valstar, Ron Rademaker, and Ludo Maat. 2005. Web-based database for facial expression analysis. In *2005 IEEE international conference on multimedia and Expo*. IEEE, 5–pp.
[42] R.W. Picard. 2003. Affective computing: challenges. *International Journal of Human-Computer Studies* 59, 1-2 (jul 2003), 55–64. https://doi.org/10.1016/S1071-5819
[43] Saranya Rajan, Poongodi Chenniappan, Somasundaram Devaraj, and Nirmala Madian. 2020. Novel deep learning model for facial expression recognition based on maximum boosted CNN and LSTM. *IET Image Processing* 14, 7 (2020), 1373–1381.
[44] M. E. L. Redondo, A. Sciutti, S. Incao, F. Rea, and R. Niewiadomski. 2021. Can Robots Impact Human Comfortability During a Live Interview?. In *Companion of the 2021 ACM/IEEE International Conference on Human-Robot Interaction*. 186–189.
[45] M. E. L. Redondo, A. Vignolo, R. Niewiadomski, F. Rea, and A. Sciutti. 2020. Can Robots Elicit Different Comfortability Levels?. In *Wagner A.R. et al. (eds) Social Robotics. ICSR 2020. Lecture Notes in Computer Science*, Vol. 12483. Springer, 664–675. https://doi.org/10.1007/978-3-030-62056-1_55
[46] Harry T Reis, Harry T Reis, Charles M Judd, et al. 2000. *Handbook of research methods in social and personality psychology*. Cambridge University Press.
[47] Judy Hanwen Shen, Agata Lapedriza, and Rosalind W Picard. 2019. Unintentional affective priming during labeling may bias labels. In *2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE, 587–593.

[48] Alan Slater and Rachel Kirby. 1998. Innate and learned perceptual abilities in the newborn infant. *Experimental Brain Research* 123, 1 (1998), 90–94.
[49] E. Thorndike. 1992. *Intelligence and Its Use.* Harper's Magazine. 227–235 pages.
[50] Shih-En Wei, Varun Ramakrishna, Takeo Kanade, and Yaser Sheikh. 2016. Convolutional pose machines. In *CVPR*.
[51] Guihua Wen, Tianyuan Chang, Huihui Li, and Lijun Jiang. 2020. Dynamic objectives learning for facial expression recognition. *IEEE Transactions on Multimedia* 22, 11 (2020), 2914–2925.
[52] Erroll Wood, Tadas Baltrusaitis, Xucong Zhang, Yusuke Sugano, Peter Robinson, and Andreas Bulling. 2015. Rendering of eyes for eye-shape registration and gaze estimation. In *Proceedings of the IEEE International Conference on Computer Vision*. 3756–3764.
[53] Wen-Jing Yan, Shan Li, Chengtao Que, Jiquan Pei, and Weihong Deng. 2020. RAF-AU database: in-the-wild facial expressions with subjective emotion judgement and objective au annotations. In *Proceedings of the Asian Conference on Computer Vision*.
[54] Lijun Yin, Xiaozhou Wei, Yi Sun, Jun Wang, and Matthew J Rosato. 2006. A 3D facial expression database for facial behavior research. In *7th international conference on automatic face and gesture recognition (FGR06)*. IEEE, 211–216.
[55] Jeffrey M Zacks and Khena M Swallow. 2007. Event segmentation. *Current directions in psychological science* 16, 2 (2007), 80–84.
[56] Amir Zadeh, Yao Chong Lim, Tadas Baltrusaitis, and Louis-Philippe Morency. 2017. Convolutional experts constrained local model for 3d facial landmark detection. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*. 2519–2528.
[57] Xing Zhang, Lijun Yin, Jeffrey F Cohn, Shaun Canavan, Michael Reale, Andy Horowitz, and Peng Liu. 2013. A high-resolution spontaneous 3d dynamic facial expression database. In *2013 10th IEEE international conference and workshops on automatic face and gesture recognition (FG)*. IEEE, 1–6.