UNIVERSITÀ
DEGLI STUDI
DI PADOVA

DIPARTIMENTO
DI INGEGNERIA
DELL'INFORMAZIONE

DIPARTIMENTO DI INGEGNERIA DELL'INFORMAZIONE

Corso di Laurea in Computer Engineering

"Detecting human engagement propensity in human-robot interaction"

Relatore: Prof. Emanuele Menegatti

Correlatore: David Tessaro

Laureando: Luca Davì

ANNO ACCADEMICO 2021 – 2022

Data di laurea 13/12/2022

Abstract

While Human-Robot Interaction (HRI) concepts were science fiction a few decades ago, many of these issues are now commonplace in modern societies and have emerged as central questions that drive studies and researches. The problem addressed in this thesis is the development of a system capable of processing images retrieved from a simple robot RGB camera stream in order to extract meaningful information related to body posture, with the ultimate objective of estimating a person's engagement propensity in scenarios involving human-robot interaction.

In HRI, particular attention has been given to the concept of engagement as it has an impact on the beginning, maintenance, and conclusion of the interaction, making it essential for natural and successful human-robot interaction. However, the studies presented here focused more on detecting the presence of engagement rather than a measurement of it, and in the ones where some measures were done, it was more related to the amount of time that engagement was detected rather than intensity or a probability that such state of mind was present.

The final objective of this thesis is to describe the system's concept, present an analysis of the problem addressed and concepts utilized, describe its implementation and the technologies used without going into great detail, and then present the results of the tests that were carried out, highlighting the strengths and weaknesses of the system.

Abstract

Mentre concetti di Human-Robot Interaction (HRI) erano fantascienza qualche decennio fa, molti di questi problemi sono ora all'ordine del giorno nella società moderna e tanto da essere trattate come questioni centrali in studi e ricerche. Il problema affrontato in questa tesi è lo sviluppo di un sistema in grado di elaborare le immagini recuperate da una semplice camera RGB di un robot, al fine di estrarre informazioni significative relative alla postura del corpo, con l'obiettivo finale di effettuare una stima della propensione all'interazione di una persona in scenari che possono prevedere un'interazione uomo-robot.

Nella HRI, particolare attenzione è stata data al concetto di "engagement" in quanto esso ha un impatto sull'inizio, il mantenimento e la conclusione di un'interazione, rendendolo essenziale per avere un'interazione uomo-robot naturale e di successo. Tuttavia, gli studi qui presentati, si sono concentrati maggiormente sulla rilevazione della presenza di "engagement" piuttosto che sulla sua misurazione, e in quelli in cui sono state effettuate alcune misurazioni, esse erano più correlate alla quantità di tempo in cui è stato rilevato "engagement" piuttosto che all'intensità o alla probabilità che tale stato d'animo fosse presente.

L'obiettivo finale di questa tesi è descrivere l'idea che sta alla base del sistema creato, presentare un'analisi del problema affrontato e dei concetti utilizzati, descrivere l'implementazione del sistema e le tecnologie utilizzate senza entrare nei dettagli, e in fine, presentare i risultati dei test che sono stati effettuati, evidenziando i punti di forza e di debolezza del sistema.

Contents

Introduction

How much confidence will individuals have in robots? What kind of connection is possible between a robot and a human? How will change our notions of what it means to be human when machines will perform human-like tasks in our midst? All of these are questions that Isaac Asimov, the man who coined the term ”robotics” in the 1940s, used as the primary unit of analysis in his stories. Many of these issues are now commonplace in contemporary societies and have emerged as central Human-Robot Interaction (HRI) research questions, whereas a few decades ago, they were science fiction. The development of robots that are able to interact with people in a variety of common settings is HRI's primary focus. As a result, human dynamics and the complexity of social environments present technical and design challenges for robotics' appearance, behavior, and sensing capabilities.

The problem addressed in this thesis is the development of a system capable of processing images available from a simple robot RGB camera stream in order to extract meaningful information related to body posture, with the ultimate goal of estimating a person's engagement propensity in scenarios involving human-robot interaction.

Engagement is a complex concept, moreover when only visual information is available. To be able to detect it, nonverbal cues are essential, therefore an introduction to nonverbal communication has been reported in the first chapter. In fact, from psychology, human-science and HRI studies, nonverbal communication appears to be a reliable source of information that can be consulted through the analysis of visual cues.

Engagement is a fundamental concept in HRI. Indeed, it is the essential ingredient in any type of natural and successful interaction, moreover when a robot is part of it. In HRI literature several studies about this topic have been published, and it is easy to notice the presence of a relationship between engagement and nonverbal cues. This has been the key element during the definition of the ground idea of the system. Alongside the limitations imposed by the problem definition, the meaningful information about body posture to be extracted has been defined, and the way to produce the final result has been proposed. Finally, tests have been described and their results discussed, highlighting the strengths and weaknesses of the system.

The project was developed during an internship at the R&D department of Omitech.

There, is developed software that defines the actions and behavior of different robots exploiting only the limited amount of information available by the producers' API. Therefore, the project's target was not just about creating something for its own sake but that could be used, directly or indirectly, to increase robot functionalities.

The thesis structure is defined as follows:

- Chapter 1 gives an introduction to nonverbal communication an HRI, describing details of concepts that constitute the idea behind the created system
- Chapter 2 reports a summary of the paper of those models exploited during the creation of the system
- Chapter 3 reports the definition of the initial problem and its analysis, highlighting the requests and limitations imposed. After that, the proposed solution is given and finally, its implementation is described
- Chapter 4 offers a variety of tests in which different situations are pictured. Trials check different aspects of the system as results behavior, consistency and execution time

Chapter 1

Nonverbal communication

Nonverbal communication is a common denominator in social life. There is hardly any domain of social experience that is not connected to it. Nonverbal communication can be defined as behavior of the face, body, or voice minus the linguistic content, in other words, everything but the words [128]. There are many ways to express nonverbal information, for example through body motion (kinesics), posture, interpersonal space (proxemics), prosody-pitch-volume-intonation of voice (paralanguage), the sense of touch (haptics), eye movement-behavior-gaze (oculesics) and timing of communication (chronemics) [195].

Nonverbal communication is often hard to be aware of in daily life, and sometimes we become conscious about it only when it goes amiss [280]. With only a quick look, people can obtain a lot of important and valid information about strangers like their feelings, thoughts, personality, sociodemographic characteristics, and much else without any apparent effort. Also, people are able to take turns smoothly in conversations, an apparently simple gesture but deriving from an amazing feat of implicitly understood coordination. From fact like those, we can think that people have predictable implicit ideas about how their own nonverbal behavior influences other people. Sensible interpersonal interaction would simply not be possible if people did not share implicit understandings of what nonverbal cues are used for and what they mean. What scholars seek to do is go beyond these implicit understandings to create a science of nonverbal communication to describe it and understand its meanings, functions, origins, and impact using empirical methods. But it is important to understand that such studies are not a panacea for all problems in interpersonal communication as if they are a kind of hieroglyphics, whose mysteries once deciphered reveal to the skilled observer the overall knowledge of human relationships.

Even if nonverbal communication started as a communication and psychology subject, it is necessary to recognize that the topic is not defined by discipline. Indeed, it is a truly cross-disciplinary subject, with connections to theory and practice in adjacent fields including sociology, anthropology, ethology, education, computer science, political science,

medicine, and many subdisciplines within psychology.

1.1 Nonverbal communication literature

Scientific research on nonverbal communication and behavior can be said to be started in 1872 with the publication of Charles Darwin's book "The Expression of the Emotions in Man and Animals". In the book, he argued that all mammals, both humans and animals, showed emotion through facial expressions, and he start wondering why facial expressions of emotions take the particular forms they do [179]. On this, Darwin attributed some facial expressions to serviceable associated habits, which are behaviors that earlier in our evolutionary history had specific and direct functions. He also continued, by questioning the reason why facial expressions persist even when they no longer serve their original purposes, proposing as answer that humans continue to make them because they have acquired communicative value throughout their evolutionary history [179]. In other words, humans utilize facial expressions as external evidence of their internal state.

Although "The Expression of the Emotions in Man and Animals" was not one of Darwin's most successful books in terms of its quality and overall impact in the field, his initial ideas started the abundance of research on the types, effects, and expressions of nonverbal communication and behavior [134].

Despite the introduction of nonverbal communication in the 1800s, the emergence of behaviorism in the 1920s paused further research on nonverbal communication [134]. The necessity of a domain for nonverbal studies occurred during the 1950s and 1960s. Its develop, was facilitated by a social and academic climate in the United States that was ready for change, a culture increasingly attracted and intrigued to visual images, a society that had adopted a focus on personal relationships, and a segment of the academic community that banded together with the goal of studying human communication. The combined effect of these factors provided an inviting climate for scholars in several different disciplines to pursue the structure and effects of nonverbal behavior in social interaction.

There have been efforts to document highlights of the multi-disciplinary history of nonverbal studies [72][173], but a comprehensive history would be too long to be described here. A more common and more manageable approach is to record the history of a particular area of nonverbal study. For example, Paul Ekman [91], Alan J. Fridlund [106], Leslie A. Zebrowitz [326], and Robert Gifford [111] provide historical benchmarks for the study of facial expressions and facial features. Historical contributions to the modern study of gestures were put forward by Adam Kendon [168][167] and Jean-Claude Schmitt [279]. John Laver [189] addressed the study of vocal quality from an historical perspective. Janet B. Bavelas and Nicole Chovil [27] provided a brief history of scholarship that analyzed the coordination of some nonverbal behavior with words, prosody, and each other. Joseph B. Walther [306] highlighted the relatively recent history of nonverbal signals mediated

by various types of technology while James P. Holoka [147] and Donald Lateiner [188] showed how some nonverbal behaviors were treated in ancient Greek and Roman literature.

1.2 What can be define as nonverbal communication?

The definition of which behaviour can be defined as nonverbal communication has provoked discussion in the past. Ekman and Wallace V. Friesen [88] argued that only those nonverbal behaviours which are intentionally made to be communicative can be intended as nonverbal communication. Paul Watzlawick et al. [307] instead, proposed a hugely different view. They argued that since all behaviour conveys information, all of them can be seen as a form of communication, for example, a passenger in a train who looks straight ahead avoiding the gaze of the other passengers can be said to be communicating just as much as if he were talking to them, since those nearby usually get the message and leave him alone [47]. Both these views have been challenged by Morton Wiener et al. [309]. They criticize the idea that all behaviour can be seen as communicative, based on the fact that it should exists a distinction between signs and communication. In Wiener et al.'s terms, information, to be regarded as nonverbal communication, needs to be shown that is both transmitted and received through nonverbal behaviour. Following this reasoning, signs, would only imply an observer making an inference or assigning significance to an event or a behaviour, instead, communication would imply a socially shared signal system or code through which an encoder makes something public, and a decoder responds in a systematically and appropriately way [47]. By this view, all behaviours are potentially informative, recalling a little what proposed by Watzlawick et al., but this is something which has to be demonstrated rather than assumed, and moreover, it has also to be shown that such information is decoded appropriately to be regarded as a form of communication. Wiener et al. also challenges the view put forward by Ekman and Friesen based on the fact that it is often difficult to establish exactly what a person does intend to communicate, as a matter of fact, once it is acknowledged that a person may be unaware, mistaken or deceitful about his intentions, there is no clue in the behaviours themselves for deciding whether or not they should be regarded as intentional communications. Also, by Peter Bull view, neither intention to communicate nor awareness of the significance of specific nonverbal cues are necessary for regarding communication as having taken place, because this may happen without any conscious intention to communicate, or indeed, even against the express intentions of the encoder. Still for Bull view, awareness of the significance of specific nonverbal cues is not necessary for communication, in the sense that neither encoder nor decoder need to be able to identify the specific nonverbal cues through which a particular message is transmitted. So, for example, people may be left with the feeling that someone was upset or angry about something without being able to specify exactly what

cues were responsible for creating that impression. Indeed, it can be argued that a great deal of nonverbal communication takes this form, and that one task of the researcher in in this field is to try and identify more precisely the cues which are responsible for creating such impressions [47].

1.3 Cues, Encoder and Decoder

A nonverbal cue can be defined as any nonverbal sign expressed, consciously or not, by a person, even if it is not understandable by someone else. In nonverbal communication, an encoder is the person that, consciously or not, express a nonverbal sign/cue. A decoder, instead, is the person able to observe that sign even though he is not aware/capable of understand it.

Encoding and decoding are two concepts that are trivial only in appearance, because in some circumstances they may redefine what can be assumed to be communication in nonverbal situations [45].

The concept expressed in section 1.2 bring to other relevant elements that must be discussed in order to have, not only a better comprehension of nonverbal communication, but also to recognize when nonverbal behaviors may be communicative and what and how much information nonverbal cues bring with them.

For Bull communication requires both encoding and decoding, but encoding may take place without decoding, while decoding may also be inaccurate [47]. This allows three different kinds of distinctions for nonverbal cues based on the decoder's ability to correctly read the situation.

Firstly, nonverbal cues may be a valuable source of information about others which, in general, is neglected. For example, this may happen when an emotion is encoded by particular nonverbal cues but is not decoded appropriately by others. This scenario has been documented by Felix Deutsch [73][75][74], concluding that the awareness of postural expression is of great value in psychoanalysis, both for the analyst in providing him with clues to psychodynamics, and for the patient in helping him to become aware of his own repressed feelings through the analyst's interpretation of the particular postures adopted. According to this view, nonverbal cues are significant not because they constitute a generalized system of communication, but as a source of valuable information which only a skilled perceiver can learn to understand through careful observation.

A second possibility is that nonverbal cues are perceived as conveying a meaning which they do not in fact possess, bringing decoders to commit what can be called as decoding error. In this case, the social significance of nonverbal cues would be quite different, they might in fact be of considerable social importance, but in a negative way, because they may lead people to make erroneous attributions about others, and possibly to act upon those mistakes. An example of this, could be the common assumption that nonverbal

cues tell us a great deal about personality, but empirical research [231] has provided little support for this belief. Others factors to be considered in this specific case are the individual differences in people's ability to decode nonverbal cues, which has been shown to be substantial [240] at the point that the extent to which nonverbal cues operate as a communication system can be assumed to vary according to the perceptiveness of the decoder.

The third possibility is that nonverbal cues may be both encoded and decoded appropriately, and that in this case their importance lies in their role as a means of communication.

The importance of the previous discussion is that the social significance of nonverbal behaviour does not necessarily lie in communication and hence the importance of the encoding/decoding distinction always needs to be considered in evaluating research on nonverbal behaviour.

1.4 Global, Innate or learned?

Nonverbal cues can be said to communicate various information as about emotion, speech, individual differences and interpersonal relationships, for this reason, their significance has to be considered in the specific social contexts [45].

Particular importance is commonly attribute to nonverbal cues in the communication of emotion. This fact, derive from the observations of Charles Darwin [69], who argued that the facial expressions of emotion constitute part of an innate, adaptive, physiological response. If the facial expressions of emotion are innate, then this would suggest that they constitute a particularly important means of communicating information about emotion. Therefore, if a person is attempting to conceal the fact that he is experiencing a particular emotion, he might not succeed in suppressing all the expressive movements associated with that particular state, ending up showing in any case, some cues about it. Also, if a person tries to convey an emotion he is not experiencing, he may fail to reproduce the spontaneous expression by omitting certain important features or by mismanaging the timing with which to show them.

Relevant evidence to the innate hypothesis can be summarized as follows. Firstly, there is the evidence from cross-cultural studies which shows that facial expressions associated with six emotions (happiness, sadness, anger, fear, disgust, surprise) are decoded in the same way by members of both literate and pre-literate cultures [90]. However, as Ekman acknowledges, the demonstration of universals in decoding does not necessarily prove that the facial expressions of emotion are inherited, it simply increases the probability that this explanation is valid [85]. The only hypothesis necessary to account for universal decoding in facial expression is that whatever is responsible for common facial expressions is constant for all mankind, thus, common inheritance is one such factor, but learning experiences common to all mankind could equally well be another.

Secondly, there is the evidence from the study of children born deaf and blind. The

ethologist Irenâus Eibl-Eibesfeldt has filmed a number of such children and claims that they show the same kinds of basic facial expressions in appropriate situational contexts as do children born without such handicaps [83]. Again, a likely explanation for these observations is that such expressions are inherited, but it is still possible that they may be learned through some form of behaviour shaping [47].

Thirdly, there is evidence from studies of non-handicapped children which shows that the facial musculature is fully formed and functional at birth. Harriet Oster and Ekman, using Ekman and Friesen's Facial Action Coding System [86], have shown that all but one of the discrete muscle actions visible in the adult can be identified in new-born infants, both full-term and premature [226]. Again, however, this does not prove that the association of particular facial expressions with particular emotions is innate. Smiling can be called a universal gesture in the sense that it is an expression which human beings are universally capable of producing, but this does not mean that it is innately associated with the emotion of happiness, nor that it has a universal meaning [47].

Thus, although the evidence is consistent with the hypothesis that certain facial expressions of emotion are innate, it is by no means conclusive. But, if the innate hypothesis is accepted as valid, then it suggests that facial expression is of particular importance in communicating information about certain emotions. However, it should be stated that this is not meant to imply that all facial expressions of emotion are innate. The learned and innate aspects of emotional expression have been neatly reconciled by Ekman in what he calls his neuro-cultural model of emotional expression [84], according to which he assumes the existence of at least six fundamental emotions with innate expressions which can be modified by the learning of what he calls display rules. These refer to norms governing the expression of emotion in different contexts and may take the form of attenuation, amplification, substitution or concealment of particular expressions.

The proposal that facial expressions of emotion may be both innate and learned has important implications for the significance which Bull attribute to facial expression in the communication of emotion [44]. For example, if this view is accepted, it would mean that no simple answer is possible to the question of the relative importance of different cues in communicating information about emotion, since it may depend on whether deliberate or spontaneous expressions have been considered. For this reason, Ekman and Friesen put forward the concept of "non-verbal leakage", for which, information about deception may be revealed more through bodily than facial cues [87]. This is based on the hypothesis that, because of the greater repertoire of facial movement, people may be more careful to control their facial movements when trying to deceive others and hence are more likely to give themselves away inadvertently through bodily movements. But, if we are comparing different types of spontaneous expression, it still seems likely that the face constitutes the prime non-verbal source of information about emotion.

An extensive literature has also been developed on individual differences both in the

encoding and decoding of nonverbal behaviour. With regard to encoding, Bull has argued that nonverbal cues may not only encode information about individual differences but that there may also be individual differences concerning the extent to which people transmit information through nonverbal cues, highlighting that, some people may transmit a great deal of information through nonverbal cues while others relatively little [46]. For example, Judith A. Hall has reviewed twenty-six studies in which comparisons were made of sex differences in encoding [127]. Nine of them showed a significant gender difference and eight of these showed that women were clearer encoders. Hence, in this sense women can be seen as more expressive, e.g., they transmit more information through nonverbal cues. Men and women also differ in the nonverbal behaviour they use as shown by Hall [126].

Individual differences in decoding nonverbal cues constitute a second important theoretical issue. An extensive research has been carried out by Robert Rosenthal et al. based on a test of decoding nonverbal cues called the Profile of Nonverbal Sensitivity (PONS) [240]. Results using the PONS show a number of significant effects due to age, sex, culture and psychopathology. The importance of these findings with regard to the communicative status of nonverbal behaviour is that although nonverbal cues may encode information about, emotion, speech or individual differences, such information may not always be accurately decoded. If certain groups of people fail to decode nonverbal cues appropriately, then the significance of those cues as a form of communication must inevitably vary according to the sensitivity of the decoders [47].

1.5 Information expressed by nonverbal communication: Emblems Illustrator and Regulators

The importance of nonverbal cues in conveying emotions has led to think to nonverbal communication as an alternative system to speech, given that it may offer a more reliable indicator of people's true feelings. This thought is reflected in the popular literature on body language, like in "Body language" by Julius Fast [101], in which it seems to be suggested that nonverbal communication represents a kind of "royal road to the unconscious", providing a vital source of information about people's "real" feelings and attitudes. From the innate hypothesis of facial expression it can certainly be argued that nonverbal cues may be a particularly important guide to people's emotions and interpersonal attitudes, but it must not be neglected the extent to which speech and nonverbal communication operate as complementary systems of communication, because, it may be the case that occurrences in which nonverbal communication conflicts with speech are the exception rather than the rule [47]. A number of researchers have in fact claimed that nonverbal behaviour is closely related to speech in terms of syntax [194], vocal stress [234] and meaning [277][276]. It has also been argued that nonverbal behaviour serves a variety of functions in relation to speech, which can be divided on the basis of a classification system proposed by Ekman

and Friesen into three main types: emblems, illustrators and regulators [88]. The term "emblems" derived from David Efron [81] to refer to those nonverbal acts which have a direct translation, such as nodding the head when meaning "Yes" or shaking the head when meaning "No", their function is communicative and explicitly recognized as such. Illustrators are movements which are directly tied to speech, and it is maintained that they facilitate communication by amplifying and elaborating the verbal content of the message. Regulators are movements which guide and control the flow of conversation, influencing both who is to speak and how much is said.

The latter type, have typically been discussed in relation to how people take turns to speak in conversation (turn-taking) and how they could try to keep it (turn-yielding) [78][80][79]. It is also possible to include under them, heading signals like greetings and farewells (also referred as access rituals by Erving Goffman [115]), which indicate a change in the amount of interaction that people have with each other's.

In relation to emblems instead, can be said that they are generally assumed to be specific to particular cultures or occupations, but some do appear to be pancultural, such as the "eyebrow flash" [82], where a person raises his eyebrows for about a sixth of a second as a greeting. Desmond Morris et al. attempted to map the geographical distribution of twenty emblems in a wide variety of locations spread across western and southern Europe and the Mediterranean, finding that some of the emblems they describe are specific only to particular cultures [214]. Kendon has also argued that, because in certain communicative contexts there may be distinct advantages in using gesture, people may also prefer to use emblems rather than speech to communicate [166]. Indeed, gesture is faster than speech, hence might be preferred where quick action is required, it is silent, hence it may be used at the same time as speech to avoid breaking in on a conversation, or to make comments on the interaction or on the participants, it is much closer to physical actions, and so it may be selected when greater impact of utterance is required, and finally, it can also be effectively received at greater distances than speech.

A number of the examples Kendon gives are of instances where emblems are used in conjunction with speech, in this case they could be said to serve the functions of illustrators. The role of illustrators as facilitators of communications is supported by many experiments. William T. Rogers, shown that the comprehension of the decoders was significantly better in a modified audio-visual condition rather than in an audio only one [259]. Reasons why this happens could be that visual information can be conveyed more easily through visual means, but also, because some gestures are like representative pictures in that they attempt to portray the visual appearance of an object, spatial relationship or bodily action ("physiographic" [81]). Jean A. Graham and Michael Argyle, in fact, tested the hypothesis that visual information is communicated more easily through hand gestures [118]. Showing that the results of the decoders were judged as significantly more accurate in the condition where gesture was permitted. Margaret G. Riseborough has

carried out a number of studies to test whether physiographic gestures facilitate communication, showing that decoders responses of what an object was, were more quickly when the description was accompanied by gesture [255]. She also argued that recall of words accompanied by gesture was significantly better than recall of words accompanied by either vague movements or no movements at all, or when noise was introduced at the same time.

Not all illustrators are physiographic in the sense described by Efron. For example, the relationship between bodily cues and vocal stress documented by Robert E. Pittenger, Charles F. Hockett and John J. Danehy suggests that body movement also supplements the information on stress communicated by changes in intonation [234]. The relationship between body movement and the syntactic and semantic structure of speech documented by Jacqueline Lindenfeld [194] and Albert E. Scheflen [277][276] would also suggest that illustrators may be useful in communicating information about the structure of speech. Efron described certain movements as "ideographic", in that they traced the logical stages or direction of a line of thought [81].

Rogers discusses a number of other possible explanations for ways in which illustrators may facilitate the comprehension of speech [259]. One possibility is that they simply increase the listener's level of attention by providing greater stimulation. Another possibility is that they create a richer bimodal sensory image which better stimulates memory processes during the decoding of speech. Rogers also suggests that illustrators may serve as a visual tracking signal for the flow of speech. An alternative hypothesis is that the prime function of illustrators is not to make the message more comprehensible, but to convey information about the speaker's emotions and attitudes, both towards the content of his own message and towards other people. Indeed, Kendon has argued that gesture does not so much "illustrate" what is being said, but adds to what is being said, conveying aspects of meaning that cannot readily be conveyed in words [165].

1.6 Nonverbal information

The term "nonverbal information" is inspired by Claude E. Shannon's mathematical information theory [282]. According to Shannon, A is a signal if its states covary with the states of a source, B. A nonverbal-information research program should consist in seeking specific covariations between the states of source and signal. In Shannon's theory, an informational link between two observable behaviors or environmental changes is a mere correlation. In Shannon's sense, when a variable (e.g., observed states of a behavior) correlates with a second variable (e.g., an environmental change), we can say that the signal carries information about the source. A signal, in Shannon's sense, is informative if the state of the signal helps to predict the state of the source. Information is contingency and correlation, but it is not causal explanation nor, most importantly, meaning.

If human behavior is approached as information, the study of nonverbal information is restricted to the findings of consistent correlations between some observable range of states from the source and some observable states of signals such as bodily postures and movements. This quantitative, probabilistic approach could be useful as a tool for predicting well defined patterns of behavior, but it does not indicate the functions and causes of such behavior. Signals do not provide, by themselves, a causal or a functional explanation of an event, just as smoke is a signal of fire but does not explain combustion.

1.7 Nonverbal communication types

Nonverbal communication cues are expressed in various ways, but mostly through, single or groups of, body parts or conveyed through the voice.

Paralanguage can express or alter the meaning of what has been said, or even, convey emotion, by using vocal related techniques such as prosody, pitch, volume and intonation.

Kinesics is the study and interpretation of body parts and of their movements as medium of nonverbal information. It is also commonly referred as body language, a term that Ray Birdwhistell, considered the founder of this area of study, neither used nor liked [68]. Indeed, body language does not have a grammar system and must be interpreted broadly instead of having an absolute meaning corresponding with a certain movement, and so, it does not meet the linguist's definition of a language. Without ending up in specific areas that will be expressed at a later time, it is easy to make examples of how much information can be retrieved from the observation of some of the most visible parts of our body: the head and torso.

When focused on the head, facial expression interpretation is an important analysis that allows the understanding of body language and the emotions expressed. In order to form an impression of a person's mood and state of mind, it analyzes multiple facial signs as the movement of the eyes, eyebrows, lips, nose and cheeks. Also, while the presence of facial signs can be interpreted as an indication of authentic emotion, an absence of it may suggest a lack of sincerity.

The body language of the head should be considered in conjunction with that of the neck. When considering it from the point of view of posture, the head should be positioned in a way which feels natural, neither stretched nor compressed. It has been shown a relationship between prolonged poor posture of the head and neck, and negative mental states [250]. When considering it in relation to motion, the neck is the basic component for a lot of common head gestures, for example nodding (considered as a sign of saying "yes" or to acknowledge a person in a respectful manner), shaking (usually interpreted as meaning "no"), and tilting (considered in different way in conjunction with eyes and context as interest, uncertainty/questioning, thinking, being suspicious) [192].

Also when focused on the torso, kinesics signs can provide information about a per-

son's state of mind. The relative fullness or shallowness of the chest, especially around the sternum, can be a key indicator of both mood and attitude, for example when fuller and positioned relatively forward, it can be interpreted as a sign of confidence, on the opposite, when it is pulled back it indicate insecurity [63]. Touching the chest can indicate different things, for example, if done with two hands, a person may want to emphasize that they are being sincere, instead rubbing the chest, especially over the heart, can be a sign of discomfort [93]. Shoulders can be another important medium of nonverbal cues as from them, particular information can be retrieved, as: confidence (back with the chest forwards), anxiety or tension (held in a raised position), depression (weak and lacking in mobility), and if a person does not know something (shrugging).

Almost every body part can express some type of nonverbal cue, and sometimes, combination or interaction of some of them can express very precise and complex signs. For these reasons is not surprising seeing the existence of a lot of sub-area of nonverbal communication focused on the analysis of some specific parts or interactions.

1.7.1 Posture

Posture is defined as the attitude assumed by the body either with support during the course of muscular activity, or as a result of the coordinated action performed by a group of muscles working to maintain the stability [237]. Posture is conventionally understood as referring to bodily positions as distinct from bodily movements, which are usually referred to as gestures.

There have been many claims for the psychological significance of posture. It provides a lot of information about a person's emotions and attitudes [73][75][74] and it can tell much about social relationships and the structure of social interaction [277][276]. It both expresses personality and constitutes a major influence on personality formation, such that manipulation of posture can be used as a valuable therapeutic technique [251][202][203].

Several are the specific information that can be obtained from posture analysis. One of the many is the posture openness. Closed posture is a posture in which parts of the body most susceptible to trauma, as throat, abdomen and genitals, are protected or concealed using other body parts, clothing, or objects [256]. Hands' back is shown, and fingers can be clenched to form a fist. Also, because it can give the impression of hiding something or resistance to closer contact, concealing the hands may be interpreted as a sign of closed posture even though the front is exposed. Closed posture has been shown to give the impression of detachment, disinterest, and hostility and usually convey unpleasant feelings [261]. On the opposite, open posture is a posture in which the vulnerable parts of the body are exposed. The head is raised, the shirt may be unbuttoned at the neck, a bag is held on the shoulder or at the side. Hands and palms are shown and relaxed, usually with palms up and fingers spread. Open posture is often perceived as communicating a friendly and positive attitude.

Other information can be understood from the inclination of the body. During conversation, a person may lean slightly toward another person or tilt slightly away from him/her. This behavior is usually unconscious. An inclination towards can be an expression of sympathy and acceptance. Inclining away can signal dislike, disapproval, or a desire to end the conversation. Decoding studies of forward lean shown that people generally perceive it as indicating a positive attitude and as more empathic than backward lean [122]. This behavior, for example, has been noticed when people were conversing with someone they liked [207], trying to persuade [122], to increase the intimacy with, [38] and when there was a good relationship with others [26].

Another source of information is the body orientation, defined as the angle of a body to another interactants. In conversation, the participants' bodies are usually turned toward each other at an angle. When a person ignores someone else, they tend to ignore or avoid contact by showing the other person their side or back. Like actions such as leaning inward or towards the other individual signal more involvement, a direct or face to face orientation communicates greater warmth and immediacy [11]. In fact, it is not a coincidence that powerful people are perceived to more directly position their body toward others [55][125][207] and that, when communicating in groups, the individual who is faced by the most people, typically has the most influence [50]. Body orientation can also be used to protect ourself from threat and vulnerability in uncertain situations such as public spaces [56].

In general, behaviors including forward leans, direct body orientation, and interaction on the same vertical plane decrease physical and psychological distance and increase immediacy [11][145].

By the fact that mood influences muscle tone, energy level, and one's internal sense of well-being, body posture often reflects a person's current state of mind [66]. Well-being affects posture by giving it a sense of energy and balance. A person's spine will be straight, the head raised, and in general the posture appears confident [225][256]. On the opposite, malaise affects posture with a sense of tiredness. A person's shoulders may droop, and the head may be bowed down or tilted to the left or right. Stress can also affect posture subconsciously by the fact that it increases the amount of muscle tension in the body. Muscle tension or rather muscular block can also indicate the will to repress certain emotion [201], for example, when someone does not want to cry, they can tighten the jaws, which suppresses tears.

1.7.2 Gesture

A gesture, in general, is a form of nonverbal or nonvocal communication in which visible bodily actions communicate particular messages, either in place of, or in conjunction with speech, exploiting movement of the hands, face, or other parts of the body. When referred to nonvocal communication, through the use of sign language, gesture can express

the same information of a spoken language [116]. From now on instead, gesture will be described only in the context of nonverbal communication.

When someone speaks, in addition to words, his gestures allow them to communicate a variety of feelings and thoughts, often along with their body language. Gesticulation and speech work independently of each other but join to provide emphasis and meaning.

Informative gestures are a passive type of gestures that mostly provide information about the person exhibiting them rather than what he is trying to communicate [179]. As they are not a part of active communication, these gestures can occur during speech, but they may also occur independently of communication [1]. Communicative gestures, instead, are gestures that are produced intentionally or not, as a way of intensifying or modifying speech produced [1].

In the context of communicative gestures, another distinction to be done is between gestures made with the hands and arms (manual), and gestures made with other parts of the body (non-manual). Examples of non-manual gestures may include head nodding and shaking, shoulder shrugging, and facial expressions, but because they have not been the primary focus of most research regarding co-speech gesture [1], they will not be processed further.

Manual gestures are most commonly broken down into four distinct categories: Symbolic (Emblematic), Deictic (Indexical), Motor (Beat), and Lexical (Iconic) [178]. The most common are the symbolic gestures, they can be used as replacement for words. These are conventional, culture-specific gestures that can be used as replacement for words and for this reason they can occur either concurrently or independently of vocal speech. Some of them are widely recognized, fixed, and have conventionalized meanings [178], while others can have a very different significance in different cultural contexts, ranging from complimentary to highly offensive [214].

Deictic gestures are gestures that consist of indicative or pointing movements, they can occur at the same time as vocal speech or in its place, and they often function in the same way as demonstrative words and pronouns such as “this” or “that” [178].

Motor or beat gestures usually consist of short, repetitive, rhythmic movements that are closely related to prosody in verbal speech. Unlike symbolic and deictic gestures, beat gestures cannot occur independently of verbal speech and do not convey semantic information. These gestures are closely coordinated with speech to keep time with the rhythm and to emphasize certain words or phrases [206].

Iconic gestures are a type of gestures full of content and can echo or process the meaning of speech occurring together. They describe aspects of spatial images, actions, people or objects [205]. Such gestures are used in conjunction with speech and tend to be universal [167].

Lexical gestures, like motor gestures, cannot occur independently of verbal speech. Their purpose is still widely contested in the literature with some arguing that they serve to

amplify or modulate the semantic content of lexical speech [167], or that it serves a cognitive purpose in aiding in lexical access and retrieval [178] or verbal working memory [112] but more recent research suggests that lexical gestures serve a primarily socio-pragmatic role [146].

1.7.3 Proxemics

Edward T. Hall, the cultural anthropologist who coined the term "proxemics" in 1963, defined it as "the interrelated observations and theories of humans use of space as a specialized elaboration of culture" [129] but, if a more explicit definition would be provide it should be that proxemics is the study of human use of space and the effects that population density has on behaviour, communication, and social interaction.

In fact, according to Hall, the study of proxemics is valuable in evaluating not only the way people interact with others in daily life, but also "the organization of space in [their] houses and buildings, and ultimately the layout of [their] towns" [123].

The value of proxemics is not only based on its use as an evaluation criterion, but can also be actively used, it has been shown that the implementation of appropriate proxemic signals improves success in monitored behavioral situations such as psychotherapy, increasing confidence of the patient towards the therapist [164].

Hall divided the interpersonal distances between people in four distinct zones: intimate (from some cm to 46cm), personal (from 46cm to 122cm), social (from 1.2m to 3.7m), public (from 3.7m to 7.6m and more).

The distance surrounding a person forms a space. Personal space is the region that surrounds a person that he considers psychologically his own. Most people value their personal space and feel discomfort, anger, or anxiety when their personal space is violated [129]. Allowing a person to enter the personal space and enter someone else's are indicators of the relationship of those people, indeed, the further a relationship with a person is, the further the zone that will be tried to be used to communicate with him will be [95].

Entering someone's personal space is normally an indication of familiarity and sometimes intimacy, however, there are situations in when space is limited between people, and this can affect them psychologically. Research on crowding shows that increasing population density has pathological effects on individuals' physiological functioning and behavior [5][53][62]. Is common in modern society, especially in crowded urban communities, to not be able to maintain our own personal space, for example when in a crowded train, elevator or street. Though it is accepted as a fact of modern life, many people find such physical proximity to be psychologically disturbing and uncomfortable [129]. In an impersonal, crowded situation like that, other nonverbal signs are commonly manifested as the tendency to avoid eye contact to try to avoid interaction with other people [7].

Another important concept in proxemics is the territory. While personal space de-

scribes the immediate space surrounding a person, territory refers to the area which a person may claim to and defend against others [211]. Scholars maintain that territoriality is partly an innate biological drive rooted in human nature [19]. Altman conceptualized three types of territories [9]. Primary, where people have executive rights to the space such as one's home (home territory). Secondary, where people interact with acquaintances in semipublic places such as a neighborhood bar (interactional territory). Public territories, where everyone has temporary access such as the beach (public territory).

Personal space is highly variable, due to cultural differences and personal preferences [292] [129]. Also, several relationships may allow for personal space to be modified, including familial ties, romantic partners, friendships and close acquaintances, where there is a greater degree of trust and personal knowledge.

1.7.4 Oculesics

Oculesics, is the study of eye movement, eye behavior, gaze, and eye-related nonverbal communication cues.

For humans, the eyes are particularly useful in establish mental and emotional states of others. The same structures that surround and protect the eyes (lids, brows, conjunctiva, lachrymal glands) have been widely implicated as social cueing mechanisms facilitating nonverbal communication [89]. Thus, the analysis of these structures in terms of social and emotional expression has been focused on the complex muscle patterning around the eyes.

The two eye behaviors receiving the most empirical and theoretical attention in the literature include eye gaze, and pupil dilation/constriction.

1.7.4.1 Eye gaze

Information from the eye region has proven particularly critical to nonverbal communication and to correctly identifying basic emotions such as sadness, fear, and anger expressions [4]. Where the eye region alone has been found to be as informative as the whole face, the former is even more important when trying to deduce complex mental states [25] or to accurately infer complex emotions when presented separately from other regions of the face [21].

Eyes analysis does not stop to that, they are capable of their own socially meaningful behaviors. Researchers who, early on did suggest that gaze might exert an influence on emotion processing, tended to agree that direct eye gaze might increase the intensity of all emotional facial displays [170][169]. However, the first studies considered important only one dimension of emotional experience, the "valence" (positive versus negative). This perspective assumes that if an effect can be shown for both a positive (e.g. joy) and a negative (e.g. anger) emotional display, then it is likely that such an effect will generalize

across all instances of specific emotions. Other meaningful dimensions along which to differentiate emotions exist, as the approach/avoidance motivational orientation, used as base for the "shared signal hypothesis" [3][2]. By this aspect, for example, anger and fear share a negative valence, but are distinguished by behavioral motivation: fight/approach, fight/avoidance. Although research related to gaze and emotion had been quite limited in the past, what had been previously done generally supports the shared signal hypothesis [21][102][92][172][265][35][270][143]. Indeed, direct eye gaze is known to signal an increased likelihood of approach and social engagement [94]. Conversely, gaze aversion is a signal of avoidance and at times considered itself an act of hiding [249][57]. Shifting of eye gaze can be also a powerful moderator of emotional distress [319].

The direction of another person's gaze can modulate the looking behavior of the decoder as well. When another person looks at you, you may reciprocate with eye contact, or you may look away. If you see this person looking off at some object in the environment, however, you are likely to follow that gaze towards the object in order to see what is being looked at [107].

Cooperative behavior can also be impacted by the perception of direct gaze. For example, simply displaying posters with pictures of eyes with direct gaze in a cafeteria reduced the amount of littering behavior that ensued [97] even without any anti-littering message. The mere reminder that one is being watched, via presentation of direct eye gaze, is sufficient enough to promote prosocial behaviors.

1.7.4.2 Pupil dilation

The pupil is the opening at the center of the iris that admits light into the eye. The constriction and dilation of the pupillary aperture is produced primarily through autonomic nervous system control exerted on the muscles of the iris, the sphincter papillae, and the dilator pupillae. These movement patterns form the basis of several optical reflexes, including the pupillary light reflex (a change in pupil diameter in response to luminance levels) and pupillary reflex dilations (pupillary responses to psychosensory stimulation).

Charles Darwin pointed out the relation of pupillary responses to autonomic nervous system activity in his book "The Expression of the Emotions in Man and Animals", noting a possible relationship between pupil dilation and fear [69].

Early work conducted by Eckhard Hess and his colleagues popularized the study of the pupillary response as feedback to social stimuli and as a signal of social responsivity [142], concluding that pupil size could be used to index level of interest in a visual stimulus. Hess subsequently extended these findings to include "bi-directional" responses, for which, pupil exhibit extreme dilation for interesting or pleasing stimuli and extreme constriction for material that is unpleasant or distasteful to the viewer [140][141]. This contention was supported in a number of other studies [144][235][208], however, a number of subsequent academic work have attempted to replicate Hess's "bi-directional" effect to no avail. The

preponderance of them found pupil dilation for emotionally engaging stimuli, regardless of valence [159].

In a more recent review concerning pupil size and mental activity, Jackson Beatty and Brennis Lucero-Wagoner concluded that the amplitude of pupillary dilation is an index of brain activity in response to the cognitive demands of memory, language processing, reasoning, and perception [29]. Further, pupil dilation is commonly listed as a component of the physiological orienting response, an alerting mechanism elicited by unexpected, novel, and significant stimuli [260]. Because of these aspects, it follows that pupillary dilation in response to psychosensory stimulation reflects attention to and analysis of the stimulus. Consequently, the amplitude of the reaction depends upon the degree of arousal that the stimulus causes.

1.8 Haptic

Haptic communication refers to the ways in which people and animals communicate and interact via the sense of touch. Touch is the most sophisticated and intimate of the five senses [49].

Touch occurs in numerous forms. Matthew J. Hertenstein noted that touch can vary in its location, frequency, duration, action, intensity, and extent [136]. As there are many ways in which one person can touch another, it is important to note how aspects such as position can affect how touch is understood and evaluated and what kind of relationship it involves [104].

Richard Heslin separated touch into five categories based on usage, function, and intensity. These categories are functional-professional, social-polite, friendship-warmth, love-intimacy, and sexually arousing touch [139]. Functional-professional or instrumental touch, the least intense or personal category, occurs in institutional settings constrained by rules of professional conduct [139]. Social-polite touch occurs in first-meetings, business, and formal occasions often in the form of a handshake [13]. This function of touch signals respect and inclusion as well as conveying some degree of equality [139]. The friendship-warmth function of touch is both the most important and the most relationally negotiated between partners. Touch in private bodily areas or excessive touch may convey sexual interest, whereas too little touch may suggest detachment or indifference and may hinder friendship or the potential for relational development. The love-intimacy touch is personal and distinctive because only people in relationships such as romantic partners, good friends, and close family members can exchange these touches. Kisses and hand-holding are examples of intimate and generally mutual touches that convey immediacy, affection, trust, and equality [48]. The most passionate, physically intimate, and private form of touch is sexually arousing touch. Mutual consent is desired when this type of touch occurs due to its stimulating, personal, and anxiety-arousing effects. Sexual arousal

can occur through many channels including words, sight, and even smell and taste, but the core of sexuality is conveyed through touch at very close interpersonal distances.

Touch conveys much more than simply warmth, it can also express specific positive and negative emotions such as anger, fear, happiness, sympathy, love, and gratitude [12][160][137]. Moreover, the accuracy with which subjects were capable of communicating the emotions were commensurate with facial and vocal displays of emotion [138].

Touch can also be used in persuasion exploiting what can be called "compliance touches". Abundant research shows touch is a potent persuasive tool in interactions with strangers in many settings. When touched appropriately, people are more willing to sign petitions [311], fill out questionnaires [120][221], positively assess service encounters [103] etc. [171][96][148][149] [291][228][163].

As for other nonverbal cues, touch has to be considered in relation with other factors because individuals vary considerably in the degree to which they like or dislike it. Considerable research has examined touch avoidance, which indicates people's liking and approach or dislike and avoidance of same-sex or opposite-sex touch [161][10][14]. Touch avoiders are less open and expressive, lower in self-esteem, but more religious than touch approachers [13]. Touch avoiders have more negative perceptions of people who touch them than do touch approachers and stay "out of touch" by utilizing larger personal distances and touching less, leading to less intimacy overall [10][14][121].

1.9 Human-Robot interaction

The study of interactions between humans and robots is known as human–robot interaction (HRI). The concept of HRI has been around for as long as the concept of robots themselves, despite the fact that it is frequently referred to as a new and developing field. Indeed, how much faith will people have in robots? What kind of connection can a robot have with a human? When machines will be doing human-like things in our midst, how do our conceptions of what it means to be human change? Are all questions that Isaac Asimov, the one who coined the term robotics in the 1940s, used as the main unit of analysis to write his stories. While these concepts were science fiction a few decades ago, many of these issues are now commonplace in modern societies and have emerged as central HRI research questions.

HRI's primary focus is on creating robots that can interact with people in a variety of common settings. This creates technical difficulties as a result of human dynamics and social environment complexities, and it also presents design challenges pertaining to the appearance, behavior, and sensing capabilities of robotics. By necessity and nature, HRI is a problem-based, multidisciplinary field. The development of robotics hardware and software, the analysis of human behavior when interacting with robots in various social contexts, the creation of the aesthetics of the robot's embodiment and behavior, and

the necessary domain knowledge for specific applications all necessitate collaboration from a variety of fields in order to create a successful interaction between human and robot. For this reason, HRI brings together researchers and practitioners from engineering, psychology, design, anthropology, sociology, and philosophy, as well as researchers and practitioners from other application and research fields.

In this multidisciplinary sense, HRI is comparable to Human-Computer interaction (HCI) and robotics. Clearly, what makes HRI unique is that the interaction of humans with social robots is at the core of this research field. These interactions usually include physically embodied robots, and their embodiment makes them inherently different from other computing technologies. HRI is concerned with the ways in which robots interact with people in the social world, whereas robotics is concerned with the creation of physical robots and the ways in which these robots manipulate the physical world. The former, on the other hand, is something that is constantly evolving, may differ depending on the location and context in which the robot is placed, and is defined by the widespread acceptance of both explicit and implicit rules. Humans might be aware of these social rules, like saying "you're welcome" when someone says "thank you". However, all of these social norms and rules are unknown to a robot and necessitate the robot designer's attention.

In HRI, this is not the only issue that needs to be taken into consideration, in fact, not only the logic and algorithm underlying the robot's behavior are crucial, but also the way the robot can perform and express its behaviors and how it looks. A robot that is embodied does not merely consist of a computer on wheels or legs. Instead, it is necessary to comprehend how to design that embodiment, both in terms of software and hardware and in terms of how it affects people and the interactions they can have with it.

In the end, HRI is more complicated than it appears, and Moravec's paradox holds true decades after it was first expressed: Anything that appears difficult to humans is relatively simple to machines, and anything that a young child can do is nearly impossible to a machine [213].

1.9.1 Noverbal communication in Human-Robot interaction

1.9.1.1 Proxemics

Humans and robots frequently share physical space. Some robots can move through the air or over the ground. In order to interact with users and objects, some of them have manipulators and arms. When designing interactions between humans and robots, it is necessary to take into account where and how these robots move in relation to people [117]. Negative reactions, as well as rejection and withdrawal from the user, will be elicited by robots that do not respect the user's personal space. Therefore, it is essential to take into consideration people's preferences and social norms regarding robot placement in relation to others when planning its placement in space. If robot designers can code the robot so

that it knows what the appropriate distance is at a given point in time and space, they can try to get people to accept the robot by keeping it at an appropriate distance and adjusting its position to provide an appropriate interaction experience.

Localization, navigation and detection

An odometer is a sensor that keeps track of how far a typical robot has traveled by its wheels. However, as the robot moves, these lose accuracy, necessitating the robot's correction of the odometry's location information. The typical approach to this problem is to let the robot create a map of its surroundings and then use information from other sensors, such as a laser range finder or camera, to cross-reference information on the robot's location and orientation from the odometry to locate itself on the map. Simultaneous localization and mapping, or SLAM, is the name given to this procedure [70][258].

Localization can assist the robot in determining the kind of space it is in, such as whether it is in the living room or the bathroom, in addition to reporting the robot's location, however, it will not reveal any information about any individuals' whereabouts in that area.

In HRI another challenge is determining where and how people are interacting with the robot. The robot will carry sensors such as two-dimensional (2D) cameras and depth cameras that enable it to identify nearby individuals in order to detect people at a short distance. The software that processes the images from the camera not only can be able to identify and follow people, but it can also be used to report the location of body parts like arms, legs, and heads. In addition, laser range finders, also known as LIght Detection And Ranging (LIDAR), are used in some methods for tracking people over longer distances.

Occasionally, a motion-capture system is utilized. Motion capture can be used to identify and locate markers (and, by extension, the people or objects they were initially attached to) by placing reflective or fiducial markers on people and objects. However, using these marker-based methods outside of a laboratory setting is difficult. Another method is to install sensors like cameras in the environment rather than on the robot itself to track people [42].

Techniques like the Dynamic Window Approach (DWA) are frequently used to prevent the robot from colliding with people or other objects [105]. The idea behind this method is that a system uses the robot's current velocity to figure out where it will be in the future while also considering whether to keep or change the robot's velocity within the limits of its actuation capability and while figuring out a future velocity that will not cause a collision.

Over longer time scales, there are techniques based on path-planning. A path-planning algorithm generates a set of waypoints or paths for the robot to follow in these methods if the goal of the robot is not immediately visible to the robot. However, when applied

to HRI, the majority of path-planning algorithms that are effective at navigating around obstacles will exhibit behavior that is inappropriate for a social context.

Socially acceptable positioning

Even though robots are able to move around without hitting anything, they often lack the ability to navigate in a way that is socially acceptable in the presence of other people. For instance, when a robot and a person are moving through a corridor in an office building, it may happen that the robot would continue to move straight down until they are just inches away from colliding, at which point it would move out of the way. Even though it would eventually avoid the person, this behavior is very different from what humans would do in a similar situation, and it can be interpreted as aggressive or confrontational. The fact that the majority of robot mapping methods only provide geometric maps with people as obstacles is the source of this issue.

How much closer to a robot do people prefer to stand? And how close should a robot get to people before it is considered rude, unsuitable, or makes them feel uneasy? Research such as [305] measured the distance at which people feel at ease when approached by robots, and according to a Hüttenrauch et al. study [152], people prefer that the robot stand at distances analogous to those considered in human proxemics.

As a result, methods for human-proxemic-based robot navigation have been developed to increase a robot's social acceptability. For instance, when a robot follows a user from behind, the robot can either follow the same trajectory as the user, or it can move directly to the user's current location, which might be a shorter and faster pathway. Gockley et al. showed that users perceive the first behavior as more natural [114]. Morales Saiki et al. developed a technique that allows a robot to navigate side by side with its user, for which they found it important for the robot to anticipate the user's future motion [212].

Perceived safety. Robots' behavior should be understandable by humans

Another thing to think about is that people's perceptions of safety may not always match what a robot thinks is safe. The behaviors of robots are typically programmed to optimize a specific task, but in order to do so, they may use movements that are difficult to comprehend and may also appear unpredictable to humans.

As a result, efforts have been made to incorporate aspects of perceived safety and comfort into path planning. For example, Sisbot et all. created a path planner for a mobile robot that plans how to achieve a specific objective while avoiding uncomfortable situations [290]. The planner considers factors like whether people are standing or sitting, as well as the possibility that the robot might surprise them by appearing from behind an obstacle.

When only a portion of the robot enters the user's personal space, it is also necessary to

plan a motion path that people will perceive as safe and comfortable. According to Kulic and Croft [181], for instance, when a robot arm is used in close proximity to a person, such as when a person and an industrial robot collaborate on a shared task, the robot must account for the socially acceptable distance when calculating a path for its end effector to achieve its specified goal (e.g., grasp an object or hand an object to a person). From a strictly functional standpoint, this could make the robot's movement inefficient, but it will result in a more favorable user evaluation of the interaction [52].

At the same time, robot motion trajectories are frequently utilized to communicate the robot's intention and objective. In order to explicitly convey information through the robot's trajectory, path-planning algorithms have been developed. A mobile robot, for instance, can indicate whether it is available for interaction by slowly moving a few meters away from a visitor [132]. In a similar way, trajectories have been used to enable cleaning robots and drones, among others, with few means of self-expression to communicate their intentions to users [295]. In HRI, when a robot hands an object to its user, this prefers a robot to behave with "legibility," which means that they can understand the robot's goal and intention [174]. Hence, researchers have developed algorithms to control a robot arm to generate legible motions while reaching a given goal. A robot could hand over an object to a person in many different ways, but the most energy-efficient one may be incomprehensible to a person, so it is better to perform a motion that is easier to be interpreted [77].

Robots should be able to understand human verbal and noverbal cues

A robot that works closely with a person should be able to comprehend how that person perceives the space around them. Consider a scenario in which two people collaborate. By saying "give me that object," one person might ask the other person to pass something to him. If there is only one object available, the referent "object" will be obvious, however, what if there are multiple items? In most cases, it is simple for people to deduce the intended referent of "object." To make the request clear, a variety of complex cues can be used, such as the person's preferences, task information, the previous context of the interaction, gaze direction, body orientation, and others. This type of interaction is defined as spatial perspective-taking, and is an important skill mostly in collaborative scenarios [302].

Spatial dynamics of initiating HRI

Every social interaction must be initiated by someone. The subsequent interaction is impacted by how someone approach each other and how this is perceived. It is generally expected that approaching behavior will benefit both parties involved in the interaction. The approacher tries to get the other person's attention, which signals interest in the person

being approached, while, initiating an interaction causes positive affect in the initiator by eliciting neural activity in reward-related brain regions [278]. In addition, starting a conversation demonstrates confidence in one's ability to have a successful social encounter and assertiveness.

Whereas this can be rather trivial for a person, a robot needs to be carefully designed to appropriately initiate an interaction. HRI has long studied robot approaching behavior. For instance, according to Nakauchi and Simmons, when a robot joins a queue, it must respect the personal space of others who are also waiting [220]. A robot's navigation mode must change from purely functional to taking into account social distance and spatial configuration when it comes into contact with people [8]. In addition, context and task depend on initiating an interaction. It have been demonstrated that if an approach is poorly planned and carried out, even a simple task as providing information about the stores in a mall will fail because of an erroneous initial approach [274][162].

The robot is not the only one that can initiate the conversation, indeed in many cases are the people themselves to try to interact with the robot. When this happens, the robot should respond precisely and with the correct timing. If it does not, the user may find the interaction to be awkward and unnatural, and they may even stop starting new interactions in the future [162].

1.9.2 Other nonverbal cues used HRI

In light of everything that has been discussed in the preceding sections, communicating effectively with a stranger may appear to be more difficult if the nonverbal communication channel is absent. This is due to the fact that while interacting, people constantly and seemingly automatically pick up on a variety of nonverbal cues. The subtleties of meaning, emotion, and intention in other people can be deduced from these cues.

Nonverbal cues that are present in human interaction have been actively utilized to enhance interactions with the robot, even in the earliest designs of social robots. They are typically used in conjunction with speech to provide additional details about the internal state or intentions of the robot. For example, one of the first social robots, Kismet, used posture cues like pulling back or leaning forward to show emotion and get people to talk to it [40]. Keepon, a minimalist social robot, demonstrates emotion and attention through gaze and reactive motion [177].

Functions of nonverbal cues in interaction

A further layer of information is added to human (and human–robot) interaction by nonverbal cues, which enable people to communicate important information between the lines.

Nonverbal communication cues, such as eye gaze, body posture, or facial muscle activity, are frequently studied as implicit indicators of affect toward a person or object, as

reported in the preceding sections. Many of the nonverbal messages that are sent are expressed automatically or even completely unconsciously. As a result, nonverbal cues are frequently regarded as unfiltered and more genuine, revealing individuals' true attitudes.

HRI relies heavily on nonverbal cues. When interacting with a robot, a person's nonverbal cues can indicate whether or not they enjoy the interaction and whether or not they like the robot. As a result, they can be used to direct the robot's behavior and serve as a measurement or cue of attitude or engagement.

The manner in which robots produce nonverbal cues may also have an impact on HRI. For instance, when the robot does not respond appropriately to people's nonverbal cues or when it makes gestures that do not match the rhythm or meaning of its speech, interaction can appear awkward.

Nonverbal cues are now widely accepted as a prerequisite for smooth and successful interaction between humans and robots, despite the fact that earlier research on HRI primarily focused on speech as the most obvious mode of communication for robots. In fact, when speaking to a robot, a person would expect that the machine would turn its head in his direction and make eye contact with him to show that it is paying attention to what he has to say. A robot that behaves in this way without speaking will make the interaction feel more natural and smoother. On the other hand, a person will notice immediately when some of this "social glue" is absent or that something is going wrong, even though it might be difficult to pinpoint exactly what is missing.

When attempting to incorporate nonverbal signals into HRI, it might be beneficial to take into account each channel of communication separately, even though humans exhibit and experience nonverbal cues in multiple modalities at once, such as sound, movement, and gaze.

Gaze and eye movement

A crucial and subtle cue for managing social interaction is gaze. People's willingness and ability to follow the conversation are also indicated by their gaze, as are interest, comprehension, and attention. In addition to their social function, gaze and eye movements facilitate functional interactions and collaboration, such as handing someone an object or pointing out the next tool needed for a task.

During an interaction, gaze can also be used to get and keep another person's attention. For instance, by looking from one person to another, the speaker might suggest whose turn it is to speak next (turn-taking management). Joint attention is a well-established aspect of gaze behavior in human interaction. Joint attention refers to interactions with partners attending to the same area or object at the same time and it is important for collaborative tasks where actors need to coordinate their activities. The timing and synchrony of gaze behavior are crucial considerations for achieving joint attention.

In several ways, joint attention has been incorporated into HRI: Imai et al. utilized it to facilitate smoother communication with individuals so that they are aware of the topic the robot is discussing, both with and without speech [153]. Joint attention has also been examined as a fundamental ability of robots, particularly humanoid robots, that want to learn from humans [275]. Finally, joint attention has been studied in interactions with children who have autism, with the aim of using robots to assist them in developing this important social skill. It is, however, still unclear whether individuals with autism who were trained to use social skills, such as performing joint attention, with robots are able to apply these skills in human–human interaction as well [257].

When used in HRI, robot gaze cues typically have the same impact as they would in human interactions. This could be because researchers have used human gaze behavior to derive models of gaze behavior for robots. They have shown that models like those, can be used to get people to take on different conversational roles like addressees, bystanders, or nonparticipants using the resulting gaze cues [216] and that it can direct who will speak next in a multiparty interaction [217]. In another study, Andrist et al. demonstrated that face-tracking movements can make a robot appear more thoughtful and intentional by engaging in mutual gaze and purposeful gaze aversions [16].

Gesture

Gesturing is perhaps the most obvious form of information transmission during an interaction, following speech.

In HRI, gesturing can also significantly improve spoken communication. The arms and hands of a robot or other body parts like its head, ears, or tail may be used to make gestures. People's perceptions and comprehension can also be affected by the shape, timing, naturalness, and smoothness of gestures [41]. Salem et al. demonstrated that the ASIMO robot used in their experiment was perceived as more anthropomorphic and likable when gestures were used in addition to speech in HRI [268]. As a result, participants expressed a greater willingness to interact with the robot later on than when the robot communicated solely through speech. This study also demonstrated that, despite having a negative impact on task performance, using gestures that were not in sync with speech resulted in even stronger positive robot evaluations.

Mimicry and Imitation

Mimicry and imitation are another aspect of nonverbal interaction that has received a lot of attention in the literature. The terms "mimicry" and "imitation" are often associated but while the former refers to the unconscious imitation of another person's behavior, the latter refers to the conscious imitation of another person's behavior.

As a largely automatic behavioral response, mimicry serves a number of important

social functions as well. One is that it indirectly signals positive affect and liking for an interaction partner. During a conversation, if two people make the same gestures or take the same posture, it usually means that they have built a positive relationship. In a similar way, you can tell that communication is not going as smoothly as it should when people's nonverbal cues are out of sync and not reflecting one another.

In the process of designing robots, various aspects of imitation and mimicry have been implemented and evaluated. There is large and growing collection of literature on robot learning by imitation, in which robots record and then replicate human actions [20]. Riek et al. developed an ape-like robot that mimicked users' head gestures, and their findings suggest this made a positive contribution to people's interactions with the robot, although these gestures were not always clear to participants [254].

If mimicry and posture from human psychology are combined, robots that are able to display certain types of behaviors (e.g., leaning in) to affect how people behave and, therefore, how they feel, can be designed. For instance, Wills et al. showed that a robot that mimicked people's facial expressions and displayed socially contingent head poses received more monetary donations than a robot that did not display such behavior [312]. As a result, mimicry and imitation can be used in HRI as conscious and unconscious social cues to improve interaction and convince people to follow the robot's recommendations.

Posture and movement

The way people move and their entire body posture also convey messages. A person's emotional state can be deduced from their postures as well as their facial expressions. These kinds of postural cues are especially important when a person's face is hidden, but they can also be an additional source of information to a person's state of mind when his facial expression can be seen. Also, in a human interaction, a person's posture can convey attention, engagement, and attraction.

Therefore, robots' bodily postures can help them express themselves even more, but also used as an alternative medium to express feelings, for example when a robot does not have expressive facial features. Beck et al. demonstrated that affective body postures can help people comprehend a robot's emotional state better [30]. Xu et al. demonstrated that humans could not only interpret the affective body postures of robots, but they could also mimic the emotions they perceived the robots to be exhibiting [320]. Designers of robots have also realized that barely perceptible micromovements can give the impression that the robot is more real [321][157][266]. These micromovements are often implemented as small, random perturbations to the robot's actuators. These lifelike animations can also be used to communicate the internal state of the robot, such as how excited the robot is by the speed or amplitude of its movement [32]. This strategy have been used successfully in non-anthropomorphized robots, such as pet-like ones, to communicate without using

human speech [64][289].

Robot perception of nonverbal cues

Standard pattern-recognition techniques are used to allow robots to perceive and identify human nonverbal cues. Typical systems use cameras, depth cameras, or sensors carried by the user to record a time series of data. Although the constant advancement of technology allows for the improvement of robotic perception capabilities, researchers also add special equipment to the robot, such as eye trackers and motion-capture systems, to provide data on nonverbal cues relevant for interaction.

Software could be written to recognize a limited number of gestures, for this reason, it is typical to train machine learning models to recognize gestures and other nonverbal cues [209].

HRI researchers use these fundamental perception techniques to estimate whether people are actually interacting with their robots. In HRI, users occasionally do not pay attention to what the robot says and signals, in contrast to typical human interaction, where the human partner is expected to be attentive and engaged. As a result, one of the most important steps in enabling robots to successfully interact with users is recognizing their engagement. Rich et al., in order to determine whether a user is engaged in interaction, they devised a method that combined back-channeling and the detection of eye contact cues [253]. Sanghvi et al. used body language and affective postures to identify engagement with a robotic game companion [271].

Chapter 2

Technologies used

2.1 Alphapose RMPE: Regional Multi-Person Pose Estimation

Multi person pose estimation in the wild is a challenging task [272][294][184][222][308] usually approached using a two-step [232][113] or a part-based [60][233][154] framework. Both of them have their advantages and disadvantages. The two-step framework first detects human bounding boxes and then estimates the pose within each box independently but, the estimation accuracy highly depends on the quality of the detected bounding boxes. The part-based framework first detects body parts independently and then assembles them to form multiple human poses, but when two or more people are too close together, body parts can be ambiguous. Also, part-based framework loses the capability to recognize body parts from a global pose view due to the mere utilization of second-order body parts dependence.

Even when this problem is simplified to a single person (single-person pose estimator SPPE), small errors in localization and recognition can cause bad results or even failures. This problem is generally inevitable especially for those methods that solely depend on human detection results. Two major problems derive from localization errors and redundant detections.

The former problem causes the SPPE to not detect body parts even when the bounding boxes are considered as correct, (e.g. with IoU > 0.5), and so, the detected human poses can be wrong. Redundant detections instead, cause the SPPE to produce a pose for each given bounding box, and so, to have redundant poses.

To address the above problems, Alphapose propose a two-step Regional Multi-person Pose Estimation (RMPE) framework that exploit three components to facilitate pose estimation in the presence of inaccurate human bounding boxes. The components are: Symmetric Spatial Transformer Network (SSTN), parametric pose Non-Maximum-Suppression (NMS), and Pose Guided Proposals Generator (PGPG).

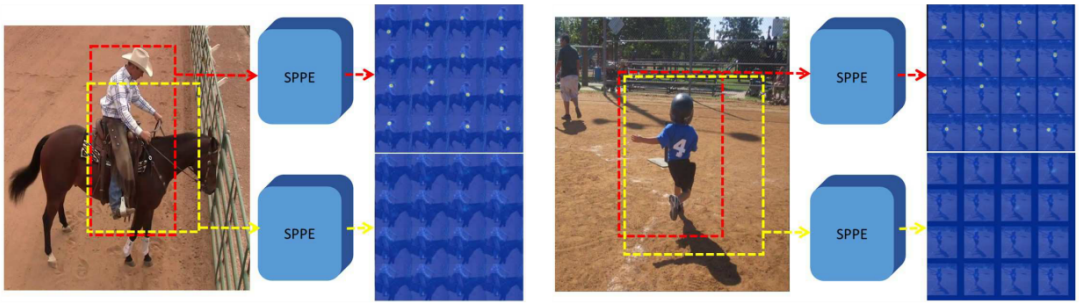


Figure 2.1: Faster-RCNN [252] detector and the SPPE Stacked Hourglass model [222] applied to show the problem of bounding box localization errors. As noted in the original paper [100], the red boxes are the ground truth bounding boxes, and the yellow ones are detected as correct bounding boxes with IoU > 0.5. The heatmaps are the outputs of SPPE corresponding to the two types of boxes. While the yellow boxes are considered as correct detections, the corresponding body parts are not detected in the heatmaps of those boxes and so, poses are not detected by the SPPE Stacked Hourglass model. Image taken from the paper [100].

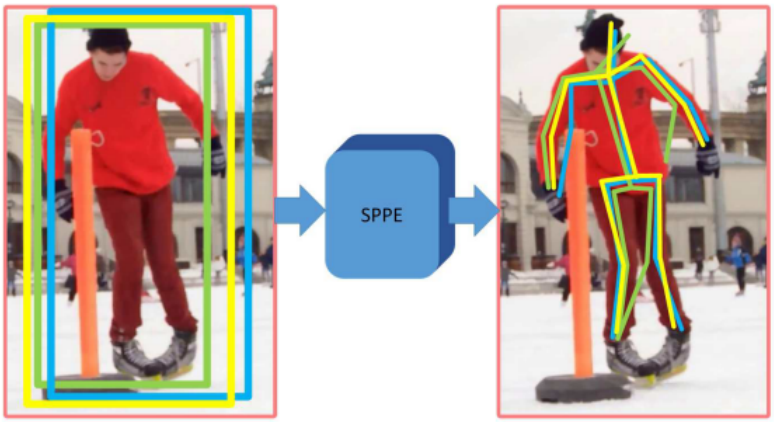


Figure 2.2: Faster-RCNN [252] detector and the SPPE Stacked Hourglass model [222] applied to show the problem of redundant human detections. As noted in the original paper [100], the left image shows the detected bounding boxes, instead the right image shows the estimated human poses. Because each bounding box is operated on independently, multiple poses are detected for a single person. Image taken from the paper [100].

The pipeline of the RMPE is illustrated in Figure 2.3. The human bounding boxes obtained by the human detector are fed into the SSTN module that consists of a Spatial Transformer Network (STN) and a Spatial De-Transformer Network (SDTN) which are attached before and after the SPPE. The STN receives human proposals and the SDTN generates pose proposals. The Parallel SPPE module acts as an extra regularizer during the training phase. Finally, the parametric Pose NMS is carried out to eliminate redundant pose estimations. Unlike traditional training, the SSTN+SPPE module is trained with images generated by the PGPG.

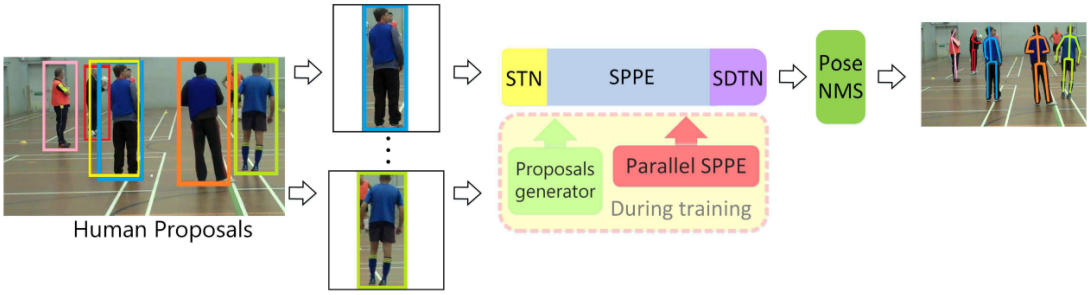


Figure 2.3: Pipeline of the Alphapose RMPE framework. Image taken from the paper [100].

2.1.1 Localization errors problem

Small translation or cropping of human proposals can significantly affect performance of SPPE [222]. The SSTN along with the parallel SPPE modules, have been introduced to the SPPE when the human proposals are imperfect. The SSTN is fine-tuned together with the SPPE.

2.1.1.1 SSTN

The STN has demonstrated excellent performance in selecting region of interests automatically, for this reason Alphapose RMPE uses it to extract high quality dominant human proposals. The STN performs a 2D affine transformation which can be expressed as:

$$\begin{pmatrix} x_i^s \\ y_i^s \end{pmatrix} = \begin{bmatrix} \theta_1 & \theta_2 & \theta_3 \end{bmatrix} \begin{pmatrix} x_i^t \\ y_i^t \\ 1 \end{pmatrix} \tag{2.1}$$

where θ_1 , θ_2 and θ_3 are vectors in $\mathbb{R}^2$. $\{x_i^s, y_i^s\}$ and $\{x_i^t, y_i^t\}$ are the coordinates before and after transformation, respectively. After SPPE the resulting pose is mapped into the original human proposal image. A SDTN is required to remap the estimated human pose back to the original image coordinate.

The SDTN computes the γ for de-transformation and generates grids based on γ:

$$\begin{pmatrix} x_i^t \\ y_i^t \end{pmatrix} = \begin{bmatrix} \gamma_1 & \gamma_2 & \gamma_3 \end{bmatrix} \begin{pmatrix} x_i^s \\ y_i^s \\ 1 \end{pmatrix} \tag{2.2}$$

Since SDTN is an inverse procedure of STN, the following can be obtained:

$$\begin{bmatrix}\gamma_1 \ \gamma_2\end{bmatrix} = \begin{bmatrix}\theta_1 \ \theta_2\end{bmatrix}^{-1} \tag{2.3}$$

$$\gamma_3 = -1 \ \times \begin{bmatrix}\gamma_1 \ \gamma_2\end{bmatrix} \ \theta_3 \tag{2.4}$$

To back propagate through SDTN, $\frac{\partial J(W,b)}{\partial \theta}$ can be derived as:

$$\frac{\partial J(W,b)}{\partial \begin{bmatrix}\theta_1 \ \theta_2\end{bmatrix}} = \frac{\partial J(W,b)}{\partial \begin{bmatrix}\gamma_1 \ \gamma_2\end{bmatrix}} \times \frac{\partial \begin{bmatrix}\gamma_1 \ \gamma_2\end{bmatrix}}{\partial \begin{bmatrix}\theta_1 \ \theta_2\end{bmatrix}} + \frac{\partial J(W,b)}{\partial \gamma_3} \times \frac{\partial \gamma_3}{\partial \begin{bmatrix}\gamma_1 \ \gamma_2\end{bmatrix}} \times \frac{\partial \begin{bmatrix}\gamma_1 \ \gamma_2\end{bmatrix}}{\partial \begin{bmatrix}\theta_1 \ \theta_2\end{bmatrix}} \tag{2.5}$$

with respect to θ_3. $\frac{\partial \begin{bmatrix}\gamma_1 \ \gamma_2\end{bmatrix}}{\partial \begin{bmatrix}\theta_1 \ \theta_2\end{bmatrix}}$ and $\frac{\partial \gamma_3}{\partial \theta_3}$ can be derived from Eqn. 2.3 and 2.4 respectively.

After having found the high quality dominant human proposal regions, the SPPE can be used for accurate pose estimation.

2.1.1.2 Parallel SPPE

The parallel SPPE branch help the STN to extract good human dominant regions during the training phase. The parallel SPPE shares the same STN with the original SPPE, but the spatial de-transformer (SDTN) is omitted. The parallel SPPE purpose is to back-propagate center-located pose errors to the STN module, indeed, the output of this branch is directly compared to labels of center located ground truth poses and because it has only to do that, all its layers are frozen. When STN extracts not center-located poses, the parallel branch will back-propagate large errors. In this way, the STN will focus on human-dominant regions.

2.1.2 Redundant detections problem

Human detectors inevitably generate redundant detections, which in turn produce redundant pose estimations. Therefore, pose NMS is required to eliminate the redundancies. The authors of Alphapose RMPE have found previous methods as [51][60] to be either not efficient or not accurate enough. For this reason they proprosed a parametric NMS method to deal with this problem.

In the following P_i indicates the i^{th} pose, each pose will be composed by m joints and each pose can be seen as a set of tuple as $\{\langle k_i^1, c_i^1\rangle, ..., \langle k_i^m, c_i^m\rangle\}$, where k_i^j and c_i^j are the j^{th} location and confidence score of joints respectively.

The parametric NMS method proposed by the author of Alphapose RMPE firstly select the most confident pose as reference, then an elimination criterion is applied to poses closed to the referenced one. This process is repeated on the remaining poses set until redundant poses are eliminated and only unique poses are reported.

The elimination criterion exploits a pose similarity function in order to eliminate the poses that are too close and too similar to each others. The function uses a pose distance metric $d(P_i, P_j|\Lambda)$ to ensure the pose similarity, and a threshold η as elimination criterion, where Λ is a parameter set of the function $d(\cdot)$. The function is therefore defined as follows:

$$f(P_i, P_j|\Lambda, \eta) = \mathbb{1}[d(P_i, P_j|\Lambda, \lambda) \leq \eta] \tag{2.6}$$

If the distance $d(\cdot)$ is smaller than η, the output of $f(\cdot)$ will be 1, which indicates that pose P_i will be eliminated due to redundancy with reference pose P_j.

The pose distance function $d_{pose}(P_i, P_j)$ is composed by two elements. Assuming that the box for P_i is B_i, the first element is $K_{sim}(P_i, P_j|\sigma_1)$ and it is defined as:

$$K_{sim}(P_i, P_j|\sigma_1) = \begin{cases} \sum_n tanh\frac{c_i^n}{\sigma_1} \cdot tanh\frac{c_j^n}{\sigma_1}, & \text{if } k_j^n \text{ is within } \mathrm{B}(k_i^n) \\ 0 & \text{otherwise} \end{cases} \tag{2.7}$$

where $\mathrm{B}(k_i^n)$ is a box centered at k_i^n, and each dimension of $\mathrm{B}(k_i^n)$ is $1/10$ of the original box B_i. The $tanh$ operation filters out poses with low-confidence scores. When two corresponding joints both have high confidence scores, the output will be close to 1. This distance softly counts the number of joints matching between poses.

The second element is $H_{sim}(P_i, P_j|\sigma_2)$. It considers the spatial distance between parts and it is defined as:

$$H_{sim}(P_i, P_j|\sigma_2) = \sum_n exp[-\frac{(k_i^n - k_j^n)^2}{\sigma_2}] \tag{2.8}$$

By combining Eqn. 2.7 and 2.8, the final distance function can be written as:

$$d(P_i, P_j|\Lambda) = K_{sim}(P_i, P_j|\sigma_1) + \lambda H_{sim}(P_i, P_j|\sigma_2) \tag{2.9}$$

where λ is a weight balancing the two distances and $\Lambda = \{\sigma_1, \sigma_2, \lambda\}$. Differently than in [60] the set of parameters are determined in a data-driven manner.

2.1.3 Pose-guided Proposals Generator

For the two-stage pose estimation, proper data augmentation is necessary to make the SSTN + SPPE module adapt to the imperfect human proposals generated by the human detector. An intuitive approach is to directly use bounding boxes generated by the human detector during the training phase. However, the human detector can only produce one bounding box for each person. The author of Alphapose RMPE implemented a proposals generator so to greatly increase that quantity. Since they already have the ground truth pose and an object detection bounding box for each person, they can generate a large sample

of training proposals with the same distribution as the output of the human detector.

They found that the distribution of the relative offset between the detected bounding box and the ground truth bounding box varies across different poses. There exists a distribution $P(\delta B|P)$, where δB is the offset between the coordinates of a bounding box generated by human detector and the coordinates of the ground truth bounding box, and P is the ground truth pose of a person.

Due to the variation of human poses, is difficult to directly learn the distribution $P(\delta B|P)$, so instead, they attempt to learn the distribution $P(\delta B|atom(P))$, where $atom(P)$ denotes the atomic pose [323] of P. To derive the atomic poses from annotations of human poses, they first aligned all poses so that their torsos have the same length. Then they used the k-means algorithm to cluster their aligned poses, and the computed cluster centers form their atomic poses. For each person instance sharing the same atomic pose a, they calculate the offsets between its ground truth bounding box and detected bounding box. The offsets are then normalized by the corresponding side-length of ground truth bounding box in that direction. After these processes, the offsets form a frequency distribution, and they fit their data to a Gaussian mixture distribution. For different atomic poses, they have different Gaussian mixture parameters.

2.2 MEBOW: Monocular Estimation of Body Orientation in the Wild

Human Body Orientation Estimation (HBOE) aims to estimate a person's orientation relative to the camera's point of view. It is important for a number of industrial applications, such as robots interacting with people and autonomous driving vehicles cruising through crowded urban areas.

HBOE can be estimated directly from image analysis or extracted from data obtained from other processes, such as human 3-D pose. Hence, it could be argued that HBOE is a simpler task than the latter and directly solvable using pose estimation models. However, HBOE warrants to be tackled as a standalone problem for three reasons. First, the 3-D pose may be difficult to infer due to poor image resolution, occlusion, or indistinguishable body parts, all of which are prevalent in in-the-wild images. Second, in certain scenarios, the orientation of the body is already sufficient to be used as the cue for downstream prediction or planning tasks. Third, the lower computational cost for the body orientation model compared to a 3-D model makes it more attractive for implementation on the device.

Although HBOE has been studied in recent years [15][24][59][108][130][131][196][219][283][324][328], it can be seen that the primary bottleneck was the lack of a large-scale, high-precision, diverse-background dataset. A robust HBOE model is presented in the paper [313], however this is not the main goal of the authors. In fact, their main goal is to fill the hole in HBOE by providing a large-scale dataset for orientation estimation from

a single in-the-wild image.

2.2.1 COCO-MEBOW

In the past, the TUD multiview pedestrians dataset [15] was the most widely used dataset for benchmarking HBOE models and today it is still used for training and evaluation in recent HBOE algorithms [15][130][131][324]. This dataset consists of 5228 images captured outdoors, many of them in grayscale. Each image contains one or more pedestrians, each of which is labeled with a bounding box and a body orientation. Body orientation labels only have eight bins (front, back, left, right, diagonal front, diagonal back diagonal left, diagonal right), which can be too coarse in certain situations. Later work [130] tried to enhance the TUD dataset by providing continuous orientation labels, each of which was averaged from the orientation labels collected from five different labelers.

There are other lesser-used datasets for HBOE. Their limitations, however, make them only suitable for HBOE under highly constrained settings but not for in-the-wild applications. For example, the 3DPes dataset [23] (1012 images, 8-bin) and CASIA gait dataset [245] (19139 frames of videos capturing 20 subjects, 6-bin) have been used in [324] and [196][248], respectively. Moreover, the human bodies in the images of these two datasets are all walking pedestrians captured from a downward viewpoint by one or a few fixed outdoor surveillance cameras. The MCGRGBD datasets [197] has a wider diversity of poses and provides depth maps in addition to the RGB images. But all its images were captured indoors and from only 11 subjects.

Human orientation can also be computed given a full 3-D pose skeleton. For this reason, human 3-D pose datasets such as the Human3.6M [156], can be converted to a body orientation dataset for HBOE research. However, due to the constraint of the motion capture system, these 3-D pose datasets often cover only indoor scenes and are sampled frames of videos for only a few subjects.

Direct prediction of body orientation from an image is valid because not only labeling a training dataset is simpler but also better performance could be achieved by directly addressing the orientation estimation problem. As supporting evidence, [110] shows that a CNN and Fisher encoding-based method taking in features extracted from 2-D images outperforms state-of-the-art methods based on 3-D information (e.g., 3-D CAD models or 3-D landmarks) for multiple object orientation estimation problems.

Given the enormous success of large-scale datasets in advancing vision research, such as ImageNet [71] for image classification, KITTI [109] for optical flow, and COCO [193] for object recognition and instance segmentation among many others, the authors of the paper [313] decided to annotate one of them to fill the lack of such a dataset for HBOE task. They presented the COCO-MEBOW (Monocular Estimation of Body Orientation in the Wild) dataset, which consists of high-precision body orientation labels for 130K human instances within 55K images from the COCO dataset [193]. This dataset uses 72

bins to partition the 360°, with each bin covering only 5°, which the authors maintained to be much more fine grained than all previous datasets while within the human cognition limit.

They chose the COCO dataset as the source of images for orientation labeling because it contains rich contextual information, and the diversity of human instances captured within it in terms of poses, lighting condition, occlusion types, and background makes it suitable for developing and evaluating models for body orientation estimation in the wild. Additionally, the COCO dataset already had bounding box labels for human instances, making it easier for body orientation labeling. As can be seen in Figure 2.5, another advantage of this dataset is that the image resolution of the labeled human instances is much more diverse than, for example, TUD dataset. This is particularly helpful for training models for practical applications in which both high- and low-resolution human instances can be captured as the distance between the camera and the subject and the weather condition can both vary.

In the making they neglected all ambiguous human instances, they labeled all suitable 133380 human instances within the total 540007 images, out of which 51836 images (associated with 127844 human instances) are used for training and 2171 images (associated with 5536 human instances) for testing.

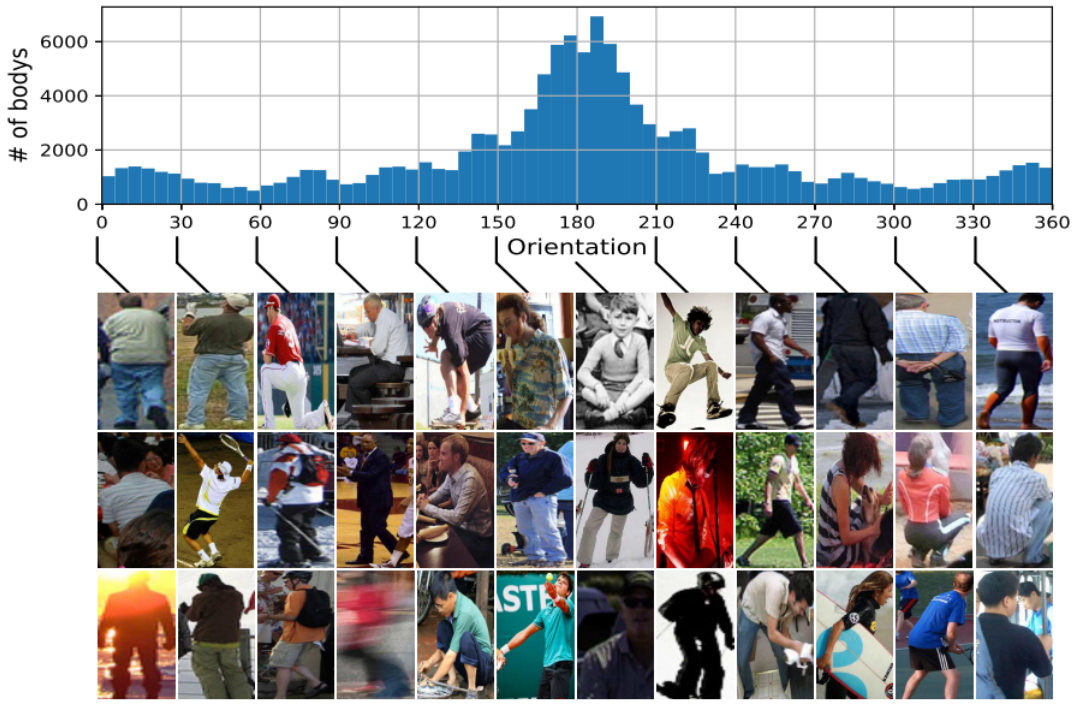


Figure 2.4: Distribution of the body orientation labels in the COCO-MEBOW dataset and examples. Image taken from the paper [313].

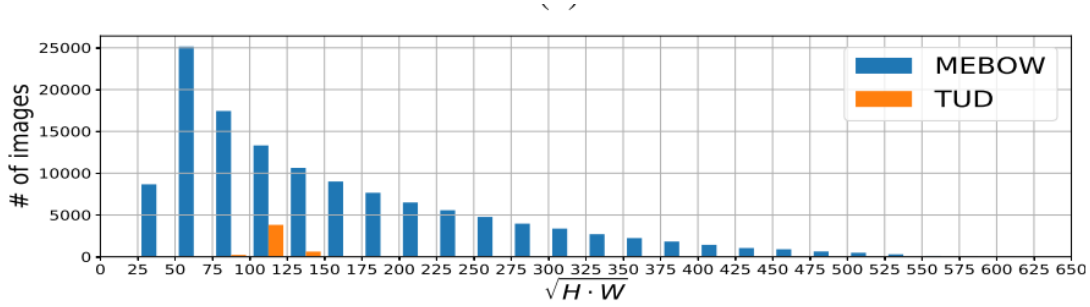


Figure 2.5: Comparison of the distribution of the captured human body instance resolution for COCO-MEBOW dataset and that for the TUD dataset [15]. The x-axis represents $\sqrt{W \cdot H}$, where W and H are the width and height of the human body instance bounding box in pixels, respectively. Image taken from the paper [313].

2.2.2 MEBOW Human Body Orientation Estimation model

Limited by the relatively small size and the coarse-grained orientation label (either 8-bin based, or 6-bin based) of existing datasets discussed above, approaches based on feature engineering and traditional classifiers such as SVM have been favored for HBOE [15][24][59][108][219][283] [328]. Deep learning-based methods [248][61] also treat HBOE as a classification problem. For example, the method in [248], given an input, it uses a 14-layer classification network to predict which bin out of eight different ones represents the orientation; the method in [61] uses a 4-layer neural network as the classification network. These methods all used simple network architecture due to the small size of the available datasets for training. And the obtained model only works for certain highly constrained environment similar to those used for training image collection. Given the continuous orientation label provided by [130] for the TUD dataset, some recent work [130][131][324] has attempted to address more fine-grained body orientation prediction. In particular, Yu et al. [324] utilized the key-points detection by another 2-D pose model as an additional cue for continuous orientation prediction. However, deep learning-based methods have been held back by the lack of a large-scale HBOE dataset.

The HBOE model architecture provided in the paper [313] can be seen in Figure 2.6. The idea behind that is that the cropped images of subjects are first processed through a backbone network as the feature extractor. The extracted features are then concatenated and processed by a few more residual layers (Head), with one fully connected layer and a softmax layer at the end.

The output are 72 neurons, $p = [p_0, p_1, \dots, p_{71}]$ $(\sum_{i=0}^{71} p_i = 1.0)$, representing the probability of every possible orientation bin being the best one to represent the body orientation of the input image. More specifically, p_i represents the probability of the body orientation θ to be within the i-th bin in Figure 2.7(b), e.g., within the range of $[i \cdot 5^\circ - 2.5^\circ, i \cdot 5^\circ + 2.5^\circ]$.

For the objective function of the model, their approach is different from previous approaches that either directly regress the orientation parameter θ (Approach 1 and 2 of

[131]) or treat the orientation estimation as a pure classification problem (Approach 3 of [131], and [130]), where each bin is a different class. Instead, they take inspiration from the heat map regression idea, which has been extremely successful in key-point estimation [222][293], and let the loss function for p be:

$$\mathcal{L} = \sum_{i=0}^{71} (p_i - \phi(i, \sigma))^2 \tag{2.10}$$

where $\phi(i, \sigma)$ is the circular Gaussian probability, as illustrated in Figure 2.7) (red curve):

$$\phi(i, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} \mathrm{e}^{-\frac{1}{2\sigma^2}(min(|i-l_{gt}|, 72-|i-l_{gt}|))^2} \tag{2.11}$$

and l_{gt} is the ground-truth orientation bin. The idea behind this is to regress a Gaussian function centered at the ground truth orientation bin so that the closer one of them is to the ground-truth orientation bin label l_{gt}, the higher the probability the model should assign to it. The validity of this idea have found foundation from the significantly eased that the learning process of the neural network had, where on the opposite, the use of standard classification loss function, such as cross entropy loss between p and the ground truth represented by one hot vector, could not converge.

As backbone model, ResNet-50 and ResNet-101 were initially considered but they observed that HRNet+Head provides much better performance in experiments. This could be explained by the fact that the HRNet and its pretrained model are also trained on COCO images and designed for a closer related task such as 2-D pose estimation.

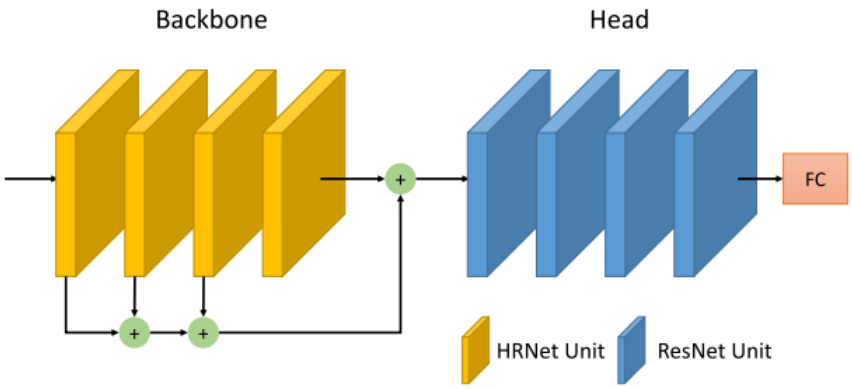


Figure 2.6: HBOE model architecture proposed in [313]. Image taken from the paper [313].

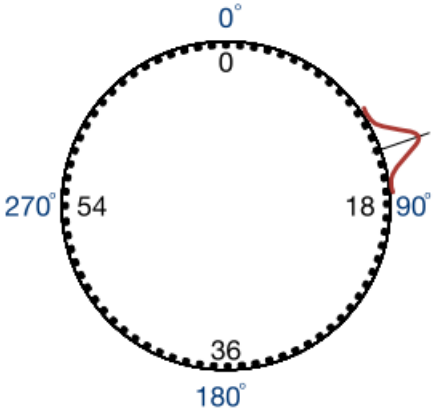


Figure 2.7: Illustration of 72 orientation bins (black ticks) and orientation loss for regressing p to the circular Gaussian target probability function. Image taken from the paper [313].

2.3 Hopenet

2.3.1 Head pose problem

In the past, head pose estimation and facial expression tracking have played an important role in driving vision technologies for non-rigid registration and 3D reconstruction and enabling new ways to manipulate multimedia content and interact with users. Major approaches to face modeling where discriminative/landmark-based approaches [273][330] and parameterized appearance models (PAM) [65][204]. More recently, due to their flexibility and robustness to occlusions and extreme pose changes, the use of modern deep learning models to directly extract 2D facial keypoints have become the dominant approach to facial expression analysis [43][331][182]. In addition, they made it possible to recover the 3D pose of the head, establishing the correspondence between the keypoints and a 3D head model and performing alignment. However, in some applications the head pose may be all that needs to be estimated and thus, the keypoint-based method may not be the best choice in that scenario.

While keypoint detectors have improved dramatically due to deep learning, head pose recovery still is a two step process with numerous opportunities for error. First, if sufficient keypoints fail to be detected, then pose recovery is impossible. Second, the accuracy of the pose estimate depends upon the quality of the 3D model of the head. Indeed, generic head models can introduce errors for any given participant, and also, the process of deforming the head model to adapt to each participant requires significant amounts of data and can be computationally expensive. Another aspect to take into account is that while it is common for deep learning based methods using keypoints to jointly predict head pose along with facial landmarks, the goal of those is to improve the accuracy of the facial landmark predictions instead to have the head pose branch sufficiently accurate on its own

[182][246][247].

A Convolutional Neural Network (CNN) architecture that directly predicts head pose has the potential to be much simpler, more accurate, and faster. While other works have addressed the direct regression of pose from images using CNNs they did not include a comprehensive set of benchmarks or leverage modern deep architectures [182][246][247] [322][227][58][119]. A brief discussion about those along with a detailed study about the weaknesses of head pose estimation from 2D landmark methods can be found in the paper of the model used in the project [263].

2.3.2 Hopenet model

Hopenet is an accurate and easy to use head pose estimation network based on a multi-loss CNN. It is trained on 300W-LP [331], a large synthetically expanded dataset, to predict intrinsic Euler angles (yaw, pitch and roll) directly from image intensities through joint binned pose classification and regression.

This model was created in order to fill a gap in the literature about non-keypoint-based head pose estimation methods and because the authors maintained that deep networks have large advantages compared to landmark-to-pose methods in that: they are not dependent on the head model chosen, the landmark detection method, the subset of points used for alignment of the head model or the optimization method used for aligning 2D to 3D points; they always output a pose prediction which is not the case for the latter method when the landmark detection method fails.

All previous work which predicted head pose using convolutional networks regressed all three Euler angles directly using a mean squared error loss. But, the authors, found that this approach does not scale well with large synthetic training data. They propose to use three separate losses, one for each angle where each loss is a combination of two components: a binned pose classification and a regression component. Any backbone network can be used and augmented with three fully connected layers which predict the angles. These three fully connected layers share the previous convolutional layers of the network.

The idea behind this approach is that by performing bin classification they use stable softmax layer and cross-entropy, thus the network robustly learns to predict the neighborhood of the pose. While, by having three cross-entropy losses, one for each Euler angle, they have three signals which are backpropagated into the network which improves learning.

In order to obtain a fine-grained predictions they compute the expectation of each output angle for the binned output. Then, they add a mean-squared error loss to the network, in order to improve fine-grained predictions. In the end, they have three final losses, one for each angle, and each is a linear combination of both the respective classification and

the regression losses. The final loss for each Euler angle is the following:

$$\mathcal{L} = H(y, \hat{y}) + \alpha \cdot MSE(y, \hat{y}) \tag{2.12}$$

Where H and MSE respectively designate the cross-entropy and Mean Squared Error loss functions, while y and $\hat{y}$ are the predicted result and the ground truth. The detailed architecture is shown in Figure 2.8.

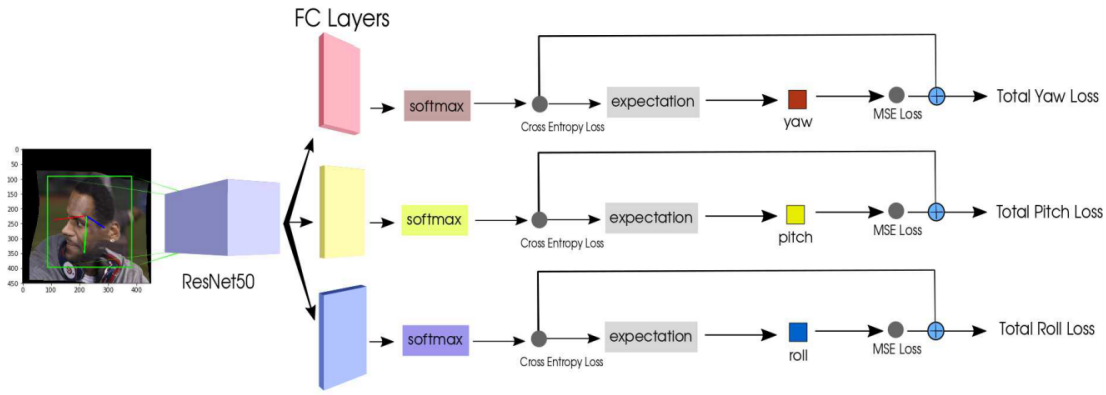


Figure 2.8: Hopenet architecture. ResNet50 [133] used as backbone. Image taken from the paper [263].

2.4 ConvNeXt

In [199] the authors review the design spaces and test the limits of what a pure Convolutional Neural Networks (CNN) can achieve. Starting from a standard Residual neural Network (ResNet) they gradually modernize it toward the design of a vision Transformer. Along the way, several key components that contribute to the performance difference are discovered, and the final outcome of this exploration leads to the definition of a family of pure CNN models called ConvNeXt. These models compete favorably with Transformers in terms of accuracy and scalability while maintaining the simplicity and efficiency of standard CNNs.

Looking back at the 2010s, the decade was marked by the monumental progress and impact of deep learning. Although the invention of back-propagation-trained CNNs dates back to the 1980s [191], it was only in late 2012 that everyone saw their true potential for visual feature learning. This awareness has led the field of visual recognition to successfully shift from engineering features to designing CNN architectures over the course of the decade. The introduction of AlexNet [180] accellerated the "ImageNet moment" [264], ushering in a new era of computer vision. Since then, the field has evolved at a rapid pace. Representative CNN like VGGNet [287], Inceptions [296], ResNe(X)t [133][318], DenseNet [151], MobileNet [150], EfficientNet [299] and RegNet [242] focused on dif-

ferent aspects of accuracy, efficiency and scalability, and popularized many useful design principles.

The dominance of CNN in computer vision was not a coincidence. Indeed, in many scenarios, the sliding window strategy is intrinsic to visual processing, particularly when working with high-resolution images. CNN have several built-in inductive biases such as translation equivariance, making them suitable for a wide variety of computer vision applications. CNN are also inherently efficient due to the fact that, when used in a sliding-window manner, the computations are shared [281].

In the meantime, neural network design for Natural Language Processing (NLP) has taken a very different path, as the Transformers have replaced recurrent neural networks to become the dominant backbone architecture. Despite the differences between language and vision domains, the two streams converged in 2020, thanks to the introduction of Vision Transformers (ViT).

ViT has made minimal changes to the original NLP Transformers, introducing no image-specific inductive bias except for the initial "patchify" layer and, with the help of larger model and dataset sizes, they can outperform standard ResNets by a significant margin. But computer vision is not limited to image classification. Without the CNN inductive biases, a vanilla ViT model faces many challenges in being adopted as a generic vision backbone. An example of this is the quadratic complexity with respect of the input size of ViT's global attention that, while it might be acceptable for image classification task as ImageNet [71], it quickly becomes intractable with higher-resolution inputs.

To fill this gap, Hierarchical Transformers uses a hybrid approach to try to behave more similarly to CNN, reintroducing the well-known "sliding window" strategy. Swin Transformer [198] represent a milestone in this direction, demonstrating for the first time that Transformers can be adopted as a generic vision backbone and achieve state-of-the-art performance across a range of computer vision tasks beyond image classification. The success of such a type of Transformer has revealed that the essence of convolution still matters and can make the difference, and that the characterizing elements of a performing architecture can also enhance others. However, the attempt to introduce ideas into architectures other than those initially used can have a cost, for example, a naive implementation of sliding window self-attention can be expensive [244] or by introducing cyclic shifting [198] the speed can be optimized but the system becomes more sophisticated in design.

From system-level comparisons such as between Swin Transformer and ResNet, it can be seen that CNN and hierarhical vision Transformers become different and similar at the same time. In fact they are both equipped with similar inductive biases but differ significantly in the training procedure and macro/micro-level architecture design.

In the following, all exploration steps done in [199] are provided but for a more exhaustive reading the paper and its appendix are suggested. They consider two model sizes

in terms of FLOPs, one is the ResNet-50 / Swin-T regime with FLOPs around $4.5 \cdot 10^9$ and the other being ResNet-200 / Swin-B regime which has FLOPs around $15.0 \cdot 10^9$. All models are trained and evaluated on ImageNet-1K [71]. Figure 2.11 shows a summary scheme with all the steps.

2.4.1 Training Techniques

The first thing that is shown in [199] is that in addition to the design of the network architecture, the training procedure also affects the final performance [31][310]. Vision Transformers have not only introduced a new set of architectural design modules and decisions, but have also introduced several vision training techniques such as the AdamW optimizer. In their study they decide to use as baseline ResNet-50/200 models, trained with a recipe close to that of DeiT's [301] and Swin Transformer's [198]. The training is extended to 300 epochs from the original 90 epochs for ResNets. They use the AdamW optimizer [200], data augmentation techniques such as Mixup [327], Cutmix [325], RandAugment [67], Random Erasing [329], and regularization schemes including Stochastic Depth [151] and Label Smoothing [297]. By itself, this enhanced training recipe increased the performance of the ResNet-50 model from 76.1% to 78.8% (+2.7%), implying that a significant portion of the performance difference between traditional CNN and vision Transformers could be due to training techniques.

2.4.2 Macro Design

In this passage, the authors of [199] analyze the Swin Transformers' macro network design. Swin Transformers follow CNN [133][288] to use a multi-stage design, where each stage has a different feature map resolution. In [199], two interesting design considerations are highlighted: the stage compute ratio, and the "stem cell" structure.

2.4.2.1 Stage compute ratio

The original design of the computation distribution across stages in ResNet was largely empirical. The heavy "res4" stage was meant to be compatible with downstream tasks such as object detection, where a detector head operates on the 14×14 feature plane. Swin-T, on the other hand, followed the same principle but with a slightly different stage compute ratio of 1:1:3:1. For larger Swin Transformers, the ratio is 1:1:9:1. Following this design, in [199] they adjust the number of blocks in each stage from (3, 4, 6, 3) in ResNet-50 to (3, 3, 9, 3), which also aligns the FLOPs with Swin-T. This improves the model accuracy from 78.8% to 79.4%. However a more optimal design is likely to exist [242][243].

2.4.2.2 Stem cell structure

Typically, the stem cell design is concerned with how the input images will be processed at the network's beginning. Due to the redundancy inherent in natural images, a common stem cell will aggressively downsample the input images to an appropriate feature map size in both standard CNN and vision Transformers. The stem cell in standard ResNet contains a 7×7 convolution layer with stride 2, followed by a max pool, which results in a $4\times$ downsampling of the input images. In vision Transformers, a more aggressive "patchify" strategy is used as the stem cell, which corresponds to a large kernel size and non-overlapping convolution. Swin Transformer uses a similar "patchify" layer, but with a smaller patch size of 4 to accommodate the architecture's multi-stage design. In [199], they replace the ResNet-style stem cell with a patchify layer implemented using a 4×4, stride 4 convolutional layer. This improves the model accuracy from 79.4% to 79.5% and suggests that the stem cell in a ResNet may be substituted with a simpler "patchify" layer similar to the one in ViT which will result in similar performance.

2.4.3 ResNeXt-ify

In this step of [199], they attempt to adopt the idea of ResNeXt [318], which has a better FLOPs / accuracy trade-off than a vanilla ResNet. The core component is grouped convolution, where the convolutional filters are separated into different groups. At a high level, the guiding principle of ResNeXt is "use more groups, expand the width". More precisely, ResNeXt employs rouped convolution for the 3×3 conv layer in a bottleneck block. As this significantly reduces the FLOPs, the network width is expanded to compensate for the capacity loss. In [199] they use depthwise convolution, a special case of grouped convolution where the number of groups equals the number of channels. The latter was observed by the author of [199] to be similar to the weighted sum operation in self-attention, which operates on a per-channel basis, for example, mixing only information in the spatial dimension. The combination of depthwise conv and 1×1 convs leads to a separation of spatial and channel mixing, a property shared by vision Transformers, where each operation either mixes information across spatial or channel dimension, but not both. The use of depthwise convolution effectively reduces the network FLOPs but also the accuracy. However, following the strategy proposed in ResNeXt, in [199] they increase the network width to the same number of channels as Swin-T's (from 64 to 96) ultimately bringing the network performance to 80.5% with increased FLOPs (5.3G).

2.4.4 Inverted Bottleneck

In [199] it is highlighted that an important design in every Transformer block is that it creates an inverted bottleneck, for example, the hidden dimension of the MLP block is four times wider than the input dimension (see Figure 2.9). But also that this transformer design

is connected to that of the inverted bottleneck with an expansion ratio of 4 used in CNN. In this passage of [199], they explore the inverted bottleneck design which configurations are illustrated in Figure 2.10 (a) to (b). Despite the increased FLOPs for the depthwise convolution layer, this change reduces the whole network FLOPs to 4.6G, due to the significant FLOPs reduction in the downsampling residual blocks' shortcut 1×1 conv layer. This results in slightly improved performance (80.5% to 80.6%) in ResNet-50 / Swin-T and even more gain (81.9% to 82.6%) in the ResNet-200 / Swin-B regime, while reducing FLOPs.

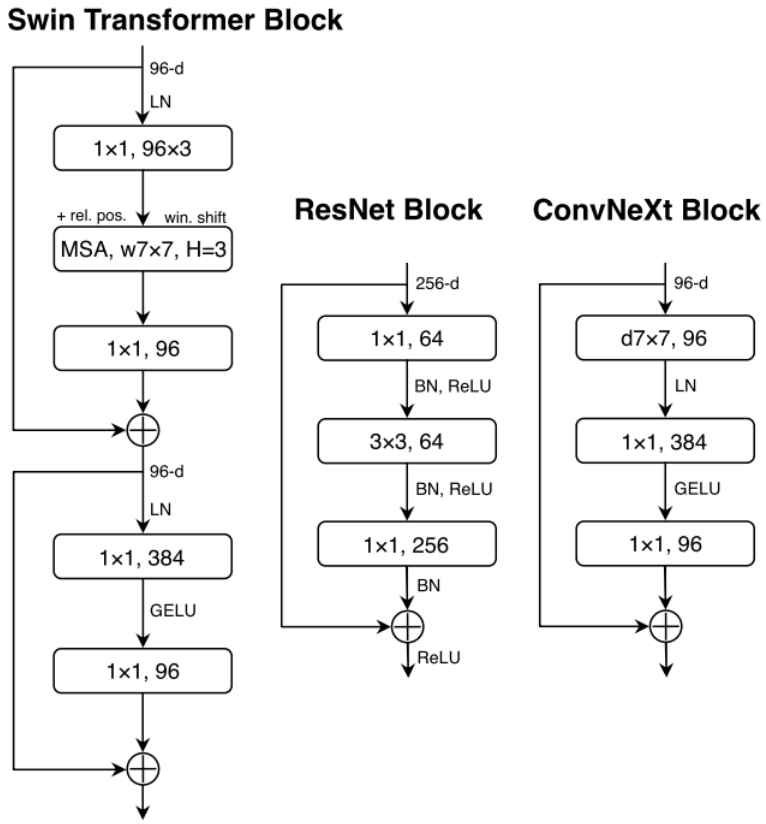


Figure 2.9: Block designs for a ResNet, a Swin Transformer, and a ConvNeXt. Image taken from the paper [199].

2.4.5 Large Kernel Sizes

In this part of the exploration, in [199] they focus on the behavior of large convolutional kernels. They point out that one of the most distinctive aspects of vision Transformers is their non-local self-attention, which allows each layer to have a global receptive field. While large kernel sizes have been used in the past with CNN [180][296], the gold standard (popularized by VGGNet [288]) is to stack small kernel-sized (3×3) conv layers, which have efficient hardware implementations on modern GPUs [190]. Although Swin Transformers reintroduced the local window to the self-attention block, the window size is at least 7×7, significantly larger than the ResNe(X)t kernel size of 3×3. For this reason,

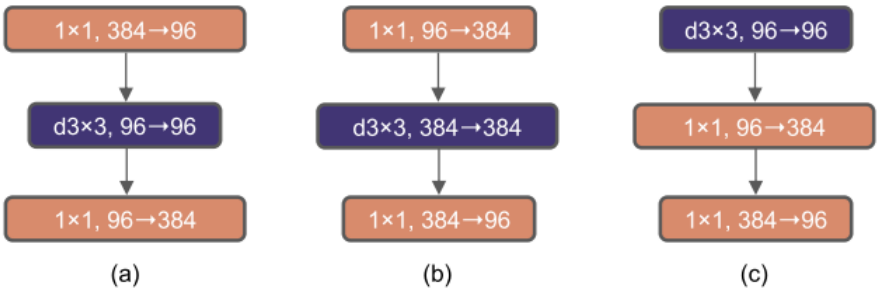


Figure 2.10: Block modifications and resulted specifications implemented by [199] in 2.4.4 step. (a) is a ResNeXt block; (b) they create an inverted bottleneck block; (c) the position of the spatial depthwise conv layer is moved up. Image taken from the paper [199].

in [199] they revisit the use of large kernel-sized convolutions for CNN.

2.4.5.1 Moving up depthwise conv layer

To explore large kernels, one prerequisite is to move up the position of the depthwise conv layer (Figure 2.10 (b) to (c)). This is a design decision also evident in Transformers: the MSA block is placed before the MLP layers. Because they have an inverted bottleneck block, complex/inefficient modules (MSA, large-kernel conv) will have fewer channels, while efficient, dense 1×1 layers will do the heavy lifting. This intermediate step reduces the FLOPs to 4.1G, resulting in a temporary performance degradation to 79.9%.

2.4.5.2 Increasing the kernel size

With all these preparations, the benefit of adopting larger kernel-sized convolutions is significant. In [199] they experimented with several kernel sizes, including 3, 5, 7, 9, and 11. The network's performance increases from 79.9% (3×3) to 80.6% (7×7), while the network's FLOPs stay roughly the same. Additionally, they note that the benefit of larger kernel sizes reaches a saturation point at 7×7 for both small and large capacity model. For this reason they use the 7×7 depthwise conv in each block.

At this point in their exploration, they have concluded their examination of network architectures on a macro scale. What can be seen at this point in the exploration made in [199], is that a significant portion of the design choices taken in a vision Transformer can be mapped to CNN instantiations.

2.4.6 Micro Design

In this section, the authors of [199] investigate several other architectural differences on a micro scale.

2.4.6.1 Replacing ReLU with GELU

In [199] they point out that a discrepancy between NLP and vision architectures are the specifics of the activation functions to be used. Numerous activation functions have been developed over time, but the Rectified Linear Unit (ReLU) [218] is still widely used in CNN due to its simplicity and efficiency. ReLU is also used as an activation function in the original Transformer paper [304]. But, the Gaussian Error Linear Unit, or GELU [135], which can be considered as a smoother variant of ReLU, is used in the most advanced Transformers, including Google's BERT [76] and OpenAI's GPT-2 [241], and, most recently, ViTs. In [199] they find that ReLU can be replaced with GELU in their CNN as well, although the accuracy stays unchanged (80.6%).

2.4.6.2 Fewer activation functions

Another thing that was noted in [199] exploration is that a small distinction between a transformer and a ResNet block is that transformers have fewer activation functions. Consider a Transformer block with key/query/value linear embedding layers, the projection layer, and two linear layers in an MLP block. There is only one activation function in the MLP block. In comparison, it is common practice to append an activation function to each convolutional layer, including the 1×1 convs. In [199] they removed all GELU layers from the residual block except for one between two 1×1 layers, replicating the style of a Transformer block (Figure 2.9). This process improves their result to 81.3%, practically matching the performance of Swin-T.

2.4.6.3 Fewer normalization layers

Still in [199], they highlight that transformer blocks usually have fewer normalization layers as well. For this reason, they removed two Batch Normalization (BN) layers, leaving only one BN layer before the conv 1×1 layers. This further increases their performance to 81.4%, already surpassing Swin-T's result.

2.4.6.4 Substituting BN with LN

Another element analyzed in [199] is Batch Normalization [155]. They maintained that BN is an essential component in CNN as it improves the convergence and reduces overfitting. However, BN also has many complexities that can have a negative effect on model performance [315]. There have been numerous attempts at developing alternative normalization [269][303][314] techniques, but BN has remained the preferred option in most vision tasks. On the other hand, the simpler Layer Normalization [22] (LN) has been used in Transformers, resulting in good performance across different application scenarios. However, directly replacing BN with LN in the original ResNet will result in subopti-

mal performance [314]. In [199], due to all the changes made to the network architecture and training techniques, using LN instead of BN they observe that their CNN model does not have any difficulties in training, on the contrary, the performance was slightly better, achieving an accuracy of 81.5%.

2.4.6.5 Separate downsampling layers

Finally in [199] they point out that in ResNet, spatial downsampling is obtained by the residual block at the beginning of each stage, using 3×3 conv with stride 2 (and 1×1 conv with stride 2 at the shortcut connection). However, in Swin Transformers, a separate downsampling layer is added between stages. For this reason they explored a similar strategy in which they use 2×2 conv layers with stride 2 for spatial downsampling. But this modification led to diverged training. To solve this problem, they conducted further investigation which showed that adding normalization layers wherever spatial resolution is changed can help stabilize training. These include several LN layers also used in Swin Transformers such as one before each downsampling layer, one after the stem, and one after the final global average pooling. This way they can improve accuracy to 82.0%, significantly exceeding Swin-T's 81.3%.

This brings them to their final model, which they have called ConvNeXt. For a more in-depth understanding but also to read about the various tests carried out by them and the relative results, it is suggested to read their paper [199].

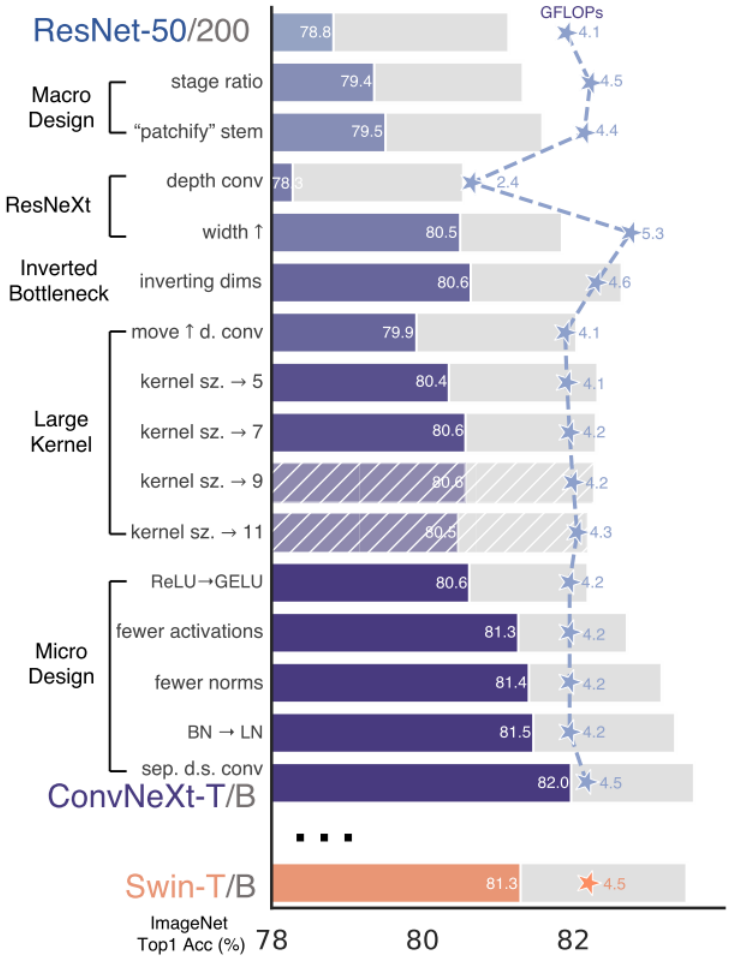


Figure 2.11: Steps in order of implementation of the gradually modernization of a standard Convolutiona Neural Network (ResNet-50) toward the design of a vision Transformer (Swin) explored in [199]. The foreground bars are model accuracies in the ResNet-50/Swin-T FLOP regime; results for the ResNet-200/Swin-B regime are shown with the gray bars. A hatched bar means the modification is not adopted. Image taken from the paper [199].

Chapter 3

Problem definition, Solution proposed and Implementation

3.1 Problem definition

The project done has been developed as solution to the problem of processing images retrieved from a simple robot RGB camera stream in order to extract meaningful information related to body posture, with the ultimate objective of estimating a person's engagement propensity in scenarios involving human-robot interaction. This level had to be expressed as a value that could be understood by humans as well as used by the robot.

Engagement with a robot partner has an impact on the beginning, maintenance, and conclusion of the interaction, making it essential for natural and successful human-robot interaction [285]. When used in HRI, this information can be exploited in different ways, for example, as a simple data to monitor the quality of certain aspects of the robot (e.g., appearance, behaviors, information expressed), or as a trigger for specific behaviors such as conversation initiation (e.g., use a more cautiously approach when moving toward a person that does not seem to want to interact, move toward the person that appears more interest in the robot) and conversation handling (e.g., change the subject if it appears that the one being discussed at the time is not in the best interest of the person with whom the robot is conversing).

In HRI studies, engagement has been viewed as both a state of mind to be achieved in an interaction where defined verbal and nonverbal cues have been exploited, and as something that must be monitored in order to manage or improve specific tasks. For instance, in [285] is presented a study on the dependence that engagement has with respect to some particular nonverbal cues such as gesture and gaze, during collaborative interaction. In [175], the nonverbal cue of a robot making eye contact is looked at to see if it would affect engagement in a HRI situation. In [267], context has been used as a characterizing element to manage definition and behavior related to engagement. In [253] instead, directed

gaze, mutual facial gaze, conversational adjacency pairs and backchannels are used as cues to recognize engagement. In [284], engagement is used to improve robot behaviors in hosting activities. In [186], nodding, laughter, verbal backchannels and eye gaze are used as social signal inputs for a real-time engagement recognition model and therefore to influence the dialogue strategy of the robot.

The system designed had to work in the most general way possible due to the fact that it had to be exploited by different social robots. In fact, nowadays more and more of them are being created, but, most of the time, they are nothing more than grey boxes with only a limited amount of controllability provided by their official API and without the possibility to add custom solution to improve their perception. Because of these factors, one of the requirements of the project was to make use of only one RGB camera, which is typically found in robots of this kind. To be able to effectively exploit information related to the robot camera stream (e.g., to modify the robot's behavior), the system had to work almost in real time. Also, a lot of the time, businesses don't make their own robots, rather, they buy them from other companies and develop software to accomplish specific jobs and commissions. In this specific project, the company had its own infrastructure to manage robots from different producers, therefore it was necessary that the system implemented was compatible with it.

Another aspect to take into account was that, because this project was done during an internship in a company, when working on a project with multiple elements like this one, it is better to design and implement each component as a standalone feature so that it can be used again in other projects. For this reason, the meaningful information that had to be extracted should have had a meaning by their self.

3.2 Solution proposed

Given requests and limitations reported above, the system had to:

- extract meaningful information related to body posture, giving priority to those that can give meaningful information even when taken alone
- estimate the engagement, or to be more precise, the level of Propensity For Interaction (PFI) that a person could have towards the robot. This level had to be expressed as a value usable both by the robot and interpretable by humans
- use information retrievable only by a single RGB camera
- work almost in real time so that information retrieved can be used to manage aspects of the robot
- be compatible with the infrastructure created by the company to manage robots from different producers

Initially, a feasibility study has been conducted in order to define if the results requested were obtainable in some way and if some algorithms or models, usable under the restriction defined, could have been already implemented. From a search in the literature, nonverbal communication immediately appeared as what should have been the key element of the project. Indeed, from psychology, human-science and also HRI studies, this element appears to be a reliable source of information which can be consulted through the analysis of visual cues [18][28][286][187][230][186][175]. This proved that meaningful information related to body posture can be found from the analysis of information retrieved from a single RGB camera.

For what concern engagement, in HRI literature several studies regarding this topic exist. In the ones in which the aspect of engagement have been examined, both explicit [33][223][262][285] and implicit measures [18][28][124][158][215][253][286][298] have been considered. However, explicit measures and questionnaires, while providing valuable hints regarding the phenomenon of interest, suffer from several limitations such as introspective ability of the user and the fact that certain aspects are implicit and automatic cognitive mechanism not accessible to conscious awareness. Instead, thanks to the careful design of experimental paradigms inspired by research in cognitive science that target specific cognitive mechanisms, objective implicit metrics could be collected and conclusions could be drawn about what cognitive processes are at stake [317][316][176]. This proved that engagement can be detected even using a single RGB camera through the detection of specific nonverbal cues expressed from the body. However, the studies analyzed, focused more on detecting the presence of engagement rather than a measurement of it, and in the ones where some measure were done, it was more related on the amount of time that engagement was detected rather than an intensity or a probability that such state of mind was present.

At this point another problem emerged. Through the visual analysis of human body, several nonverbal cues can be detected and therefore meaningful information can be extracted, also engagement can be detected by being able to determine the presence of specific nonverbal cues. However, the system has to provide results almost in real time. Then, what should be the cues to detected in order to satisfy requirements and limitations?

Engagement is a complex concept moreover when thought from a nonverbal communication point of view. From what it can be seen in Chapter 1, not only complex state of mind like this one can be expressed in different ways, but the single element that may indicate the presence or absence of it should be considered together. Another aspect to consider is that I was and I still am not an expert in this field nor there was the possibility to consult someone like that. Therefore, in order to achieve something, it was necessary to analyze this concept from an engineering perspective. Firstly, it was necessarily to simplify the concept of engagement to take into account. Considering what was expressed in other studies cited before, in this project, engagement has been considered as the presence

of specific nonverbal cues that are supposed to express willingness or even only propensity of a person for an interaction of any type, moreover with the robot recording the situation. From what expressed in Chapter 1 and in [229], nonverbal cues that may indicate this particular state of mind could be related to: gesture, body posture, body orientation, proxemics and even paralanguage. Considering the limitation to use only the RGB camera, paralanguage could not be exploited. Also most of proxemics related cues could not be detected due to the lack of depth information, and the ones that could have been estimated, would have imposed heavy bounds on the working environment or would have been derived from not reliable measurements. Gestures have also been discarded, at least in this version of the project, because of the level of complexity that they would have added (e.g., multiple tracking). Also, because single image analysis (in that case for face) were already implemented by the company, not using gestures would have simplified the adaptation of the system with the their infrastructure. Unlike the others, body posture and orientation seemed suitable for the project, but while the latter have a well defined meaning, body posture still was a too big group of nonverbal cues to be exploited as they were. For this reason, always referring to Chapter 1 and [229], and also by what have been found during the feasibility study, it was decided to consider only a sub-category of it, that is, those cues related with the openness of the body. In the end, considering what written until now and the possible meaning that such information can have when considered alone and together, it was decided to estimate:

- body pose as a set of keypoints
- body orientation with respect to the robot
- head orientation with respect to the robot
- body pose type:
  - one between: open, closed
  - one between: upstanding, sitting, lying
- hand pose type:
  - one between: open, closed
  - one between: palm, back

and that the single results would have been, in the end, aggregated in order to define a score related to the level of engagement estimated.

3.3 Solution Implementation

The system structure is reported in Figure 3.1. Media hub is a component of the infrastructure developed by the company, it manages all the media information produced by the

robot and other components, in this case it will be the one that requests the execution of the system developed in this project. Components that compose the company infrastructure exploit the websocket communication protocol to communicate with each other, for this reason the system implemented in this project will use it as well.

The Message Manager has the task of managing input and output messages. Each time that a component wants to fulfil a request, if it was not done yet, it has to establish a websocket connection with the system and send a request message. The Message Manager will process the input message and send another one to the Task Manager containing all the information necessary to it to execute the correct pipeline. The Task Manager will manage the requests to the other component based on the message received by the Message Manager, also, it will forward to the Message Manager components results whenever they are ready. As pictured in Figure 3.1, all other components are independent to each other. In fact, each one of them has been developed as an independent websocket component. By doing so, each of those can be used alone if the necessary input is given. This allows to use, test, upgrade and add components without changing the backbone structure of the whole system. Additionally, thanks to that, it was possible to define the system's request input message so that it must explicit the desired information to include in the response message. In this way if only some components' output has been requested, not all the pipeline and elements must be executed. Another thing exploitable, thanks to that level of independence, was the asynchronicity in Python, that allowed the execution of other parts of the code when an asynchronous command is executed and a result is waited, permitting in this way, to send each output whenever it is ready. The possible outputs requestable are: body pose keypoints, body orientation, head orientation, body pose type (upstanding, sitting, lying and open, closed), hands pose type (palm, back and open, closed). Despite all of this, the pipeline is not completely parallelized. In fact, some components require other elements' results as input. As it can be seen in Figure 3.1, components of different groups have different connections colors. Only members of the same group can work asynchronously, also the order of execution is expressed in the legend of the Figure 3.1. If only one part of a group has been requested, only the requested output will be included in the response message, however, all parts in the previous pipeline step will still be executed.

The system makes use of a variety of models that must be loaded from memory. To avoid hardcoded path references to models, the system makes use of a configuration file in which their path and each component settings must be declared. Also, because the system must work in the most generic way possible, a path to the file that contains the camera matrix and the distortion coefficients can be specified in the configuration. During the preprocess, those data will be used to undistort the input image. The path to the configuration file must be specified in a specific environment variable for it to be located. To make the deployment easier a docker image has been built successfully starting from a Nvidia container.

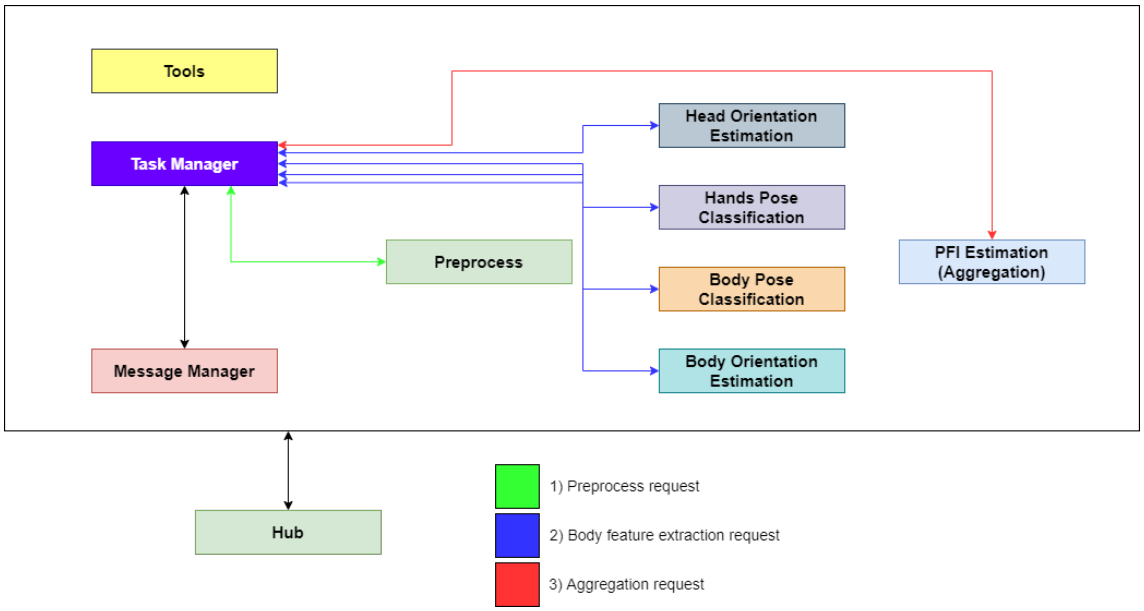


Figure 3.1: System schema

3.3.1 Components implementation

3.3.1.1 Input message

An acceptable input message is formatted as:

- client_id: id of the machine that has made the request
- flow_id: contains the sequence of actions/routes done
- encoded_image: input image encoded in base64 format
- actions: list of requested action
- camera_calibration: flag to enable the image undistortion

3.3.1.2 Preprocess

Body pose is generically estimated as a set of keypoints with a predefined order. Because of the flexible usage of its results, it could be used as the core of several applications (e.g., gesture recognition, body tracking, etc.). However in this project, it plays a minor role as it is only used for some spatial/geometry based checks and to locate people and their heads and hands. The reason why it is utilized in this manner is that, considering the study done in [263], keypoints seems not to be a reliable source of information to use directly as input for other estimations. However, this approach has been used because its results, taken alone, can be exploited easily in other projects, and to avoid employing multiple deep learning models to detect all body parts of interest (e.g., in this project 3

different detectors would have been used). Additionally, by doing so, other body parts can be detected in an easy way opening the system to further development.

The preprocess block has the role of checking if something is wrong with the input image passed in input, undistorting it if it has been requested and extracting from it, if there are any people, meaningful information about body pose.

The first thing that this component does is to check if the configuration file can be found by following the path specified in its environment variable. If it was not possible to parse the file, the execution will be stopped and an error message will be sent as response. The image is then decoded from base64 and a simple check is exploited to define whether it can be used. This simple check consists of: trying to load the image using the Opencv [39] package; extracting its edges using canny edge detector [54]; computing a pixel ratio defined as:

$$pixel\_ratio = \frac{\sum_{i=0}^{h\_edge} \sum_{j=0}^{h\_edge} p\_edge_{i,j}}{h\_edge \cdot w\_edge \cdot 255} \tag{3.1}$$

where p_edge, h_edge, w_edge are respectively the pixels, height and weight of the canny edge detector resulting image. If this ration is smaller than a fixed threshold it means that the image is too homogeneous (e.g., blurred images, camera malfunctions, something occluded the camera) and therefore, it is very likely that no information can be extracted. In order to improve the execution efficiency, whenever an image of this type is passed, the execution is interrupted and a specific error is sent as response.

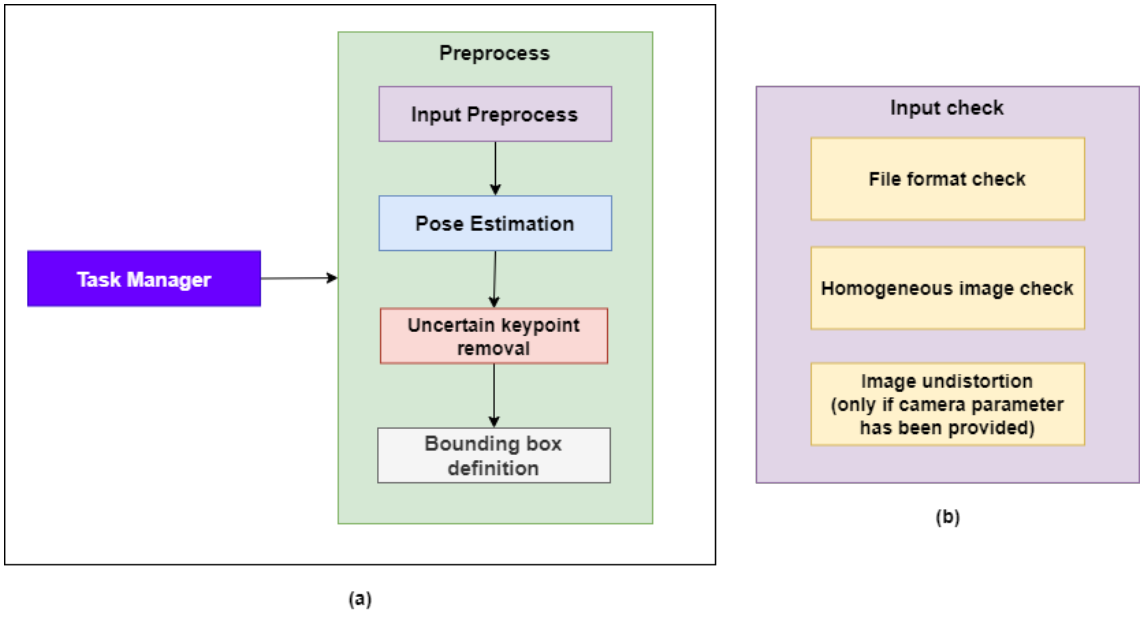


Figure 3.2: (a) Preprocess pipeline. (b) Checks done on the input message

After this, if the camera_calibration flag has been declared as true, the image is undistorted exploiting OpenCV library and its related camera calibration and undistortion func-

Figure 3.3: (a) Grey-scale converted input image. (b) Canny edge detector applied to (a) result. (c) Blurred version of (a). (d) Canny edge detector applied to (c)

tions.

Figure 3.4: (a) Input image. (b) Undistorted version of (a)

Before requesting the execution of the body pose estimator, a rapid check on the action requested is done in order to verify which components are present in it and the correctness of the request itself. In fact, it is possible that some of the requested actions are misspelled, do not exist, or have not been specified at all. If no actions have been defined correctly, the execution will be interrupted and a specific error will be sent as response, otherwise the error message will be sent along with results of correctly requested components. After this, a request to the body pose estimation component is done.

In the body pose estimation component, a version of Alphapose API slightly modified by me is used to estimate people's pose as a set of keypoints. Specifically, the 68-keypoint model trained on halpe dataset [99] has been utilized due to its balance between performance and result finesse. As result, a set of keypoints (defined by image coordinates) and

confidences are given for each person detected.

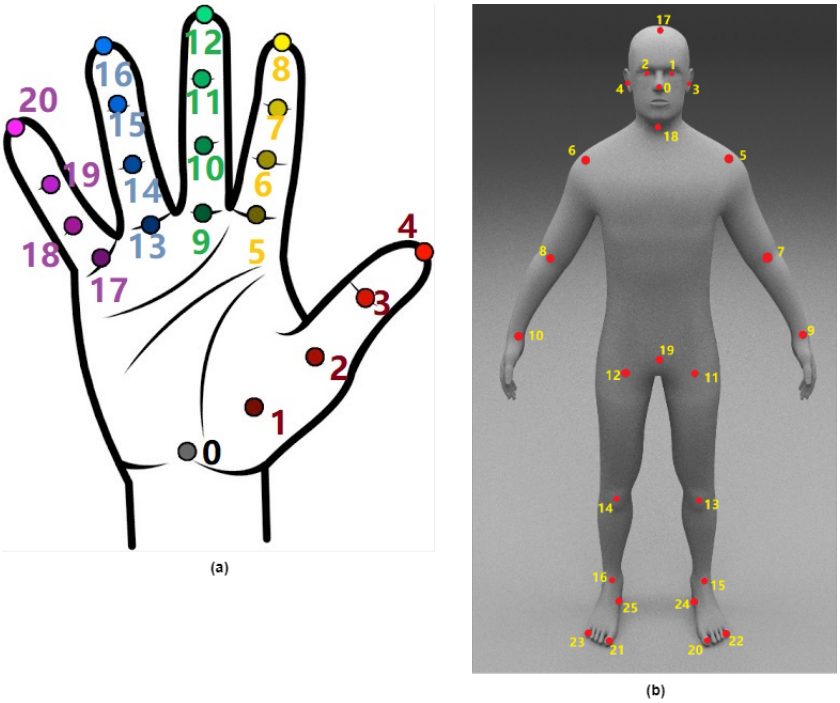


Figure 3.5: (a) Input image. (b) Undistorted version of (a)

Figure 3.6: Two example of the Alphapose result. Images taken from COCO dataset

If at least one person has been detected, keypoints with a confidence smaller than some fixed thresholds are "removed" by replacing their values and confidences with fixed ones (full body, head and hands have different thresholds). Then, full body, head, left and right hands bounding boxes are defined. In this process, firstly a height and width padding are computed as in algorithm 12:

The height padding is defined as the distance found as the Frobenius norm of left or

Algorithm 1: Height padding

if $conf[s\_l] > p\_conf\_thr$ ***and*** $conf[hip\_l] > p\_conf\_thr$ **then**
  $h\_pad = round(norm([kp[hip\_l][x], kp[hip\_l][y]] - [kp[s\_l][x], kp[s\_l][y]]))$
else if $conf[s\_r] > p\_conf\_thr$ ***and*** $conf[hip\_r] > p\_conf\_thr$ **then**
  $h\_pad = round(norm([kp[hip\_r][x], kp[hip\_r][y]] - [kp[s\_r][x], kp[s\_r][y]]))$
else
  if $w\_image > h\_image$ **then**
    $h\_pad = image\_ratio \cdot h\_image$
  else
    $h\_pad = image\_ratio \cdot w\_image$
  end
end

Algorithm 2: Width padding

if $conf[s\_l] > p\_conf\_thr$ ***and*** $conf[s\_r] > p\_conf\_thr$ **then**
  $w\_pad = round(norm([kp[s\_l][x], kp[s\_l][y]] - [kp[s\_r][x], kp[s\_r][y]]))$
else
  if $w\_image > h\_image$ **then**
    $w\_pad = image\_ratio \cdot h\_image$
  else
    $w\_pad = image\_ratio \cdot w\_image$
  end
end

right shoulder-hip keypoints coordinates. If both keypoints of a side have been flagged as "removed" their couple will not be considered. If both sides cannot be considered, a fraction of the smaller image dimension will be used instead. The width padding is defined analogously as the previous one but considering as the only couple, the left and right shoulders keypoints coordinates. By doing so, the padding will be related with the body dimension of each person. A fraction of these paddings will be used to adjust the dimensions of the bounding boxes.

The method for finding each bounding box is the same, but keypoints and the amount of padding used will vary depending on their type. Each bounding box will be initially defined as the rectangle which left-upper corner is the keypoint with minimum coordinates values among the ones considered for that box type and the right-lower corner as the one with maximum coordinates values ("removed" keypoints will not be considered). The sizes of the boxes are eventually increased by a fraction of the padding found before. To prevent this increase from causing the boxes to exceed the size of the image, boxes dimensions are checked in relation to the ones of the image and modified so to not exceed them.

In the end, the number of people detected, keypoints-confidences and bounding boxes are sent back as response along with error descriptions if any occurred.

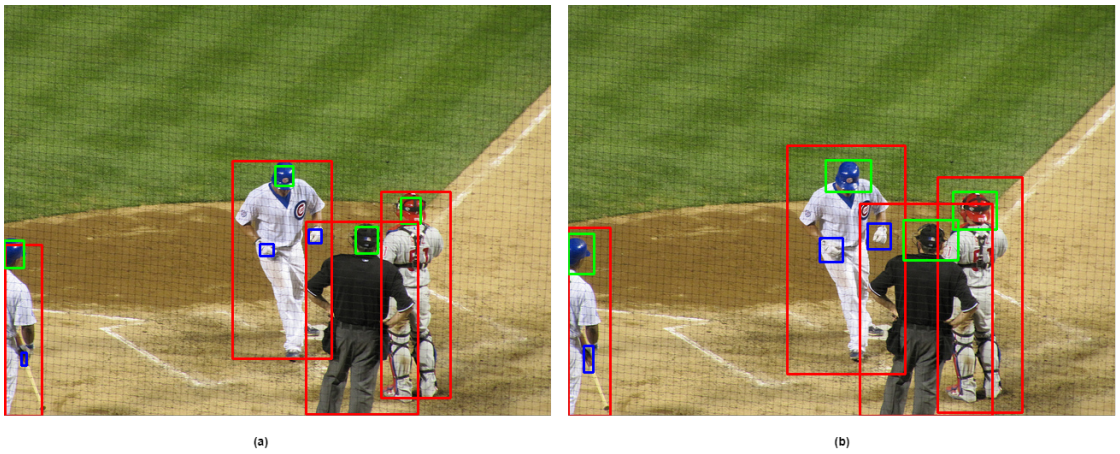


Figure 3.7: Bounding box definition. (a) before padding. (b) after padding

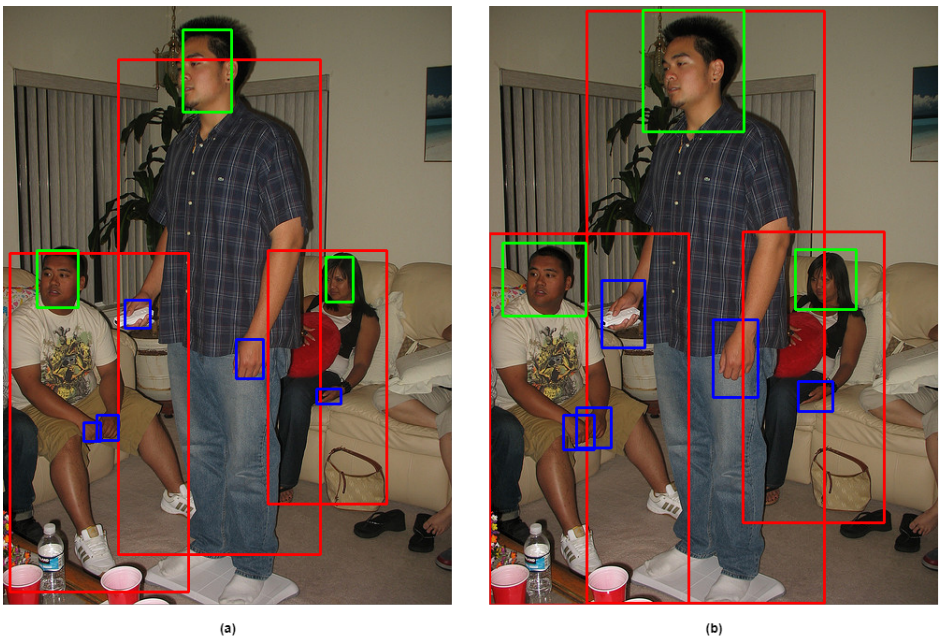


Figure 3.8: Bounding box definition. (a) before padding. (b) after padding

3.3.1.3 Head Orientation

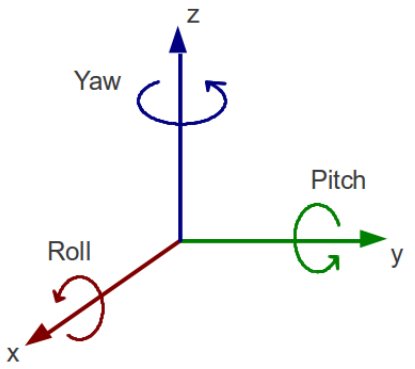
Oculesics have great impact in nonverbal communication, indeed in 1.7.4 has been reported that the eyes are particularly useful in establish mental and emotional states of others and that they have been widely implicated as social cueing mechanisms facilitating nonverbal communication. However, a direct visual analysis of eyes is not always feasible. Indeed, eyes are very small and even smaller are the part of them that may express features (e.g., direction and pupil dilation). In relation to the project, the robot would have to be placed very close to the person that have to be analyzed, reducing significantly the number of possible uses of the system. Additionally, this will impose a strong starting proxemic cue related to the person itself. This is not detectable by the system and therefore, it can cause unpredictable results. For these reasons, a simplification of that cue has

been considered for the project. Having a greater dimension, typically follows the eye or even just the direction of interest, head orientation can be detected in a more reliable way and in more general situations, therefore, it has been considered for the project.

Head orientation block has the role to estimate the yaw, pith, and roll of heads detected by the preprocess component. To do so, it requires the input image, the keypoints and the bounding boxes previously extracted.

The estimation of those angles is entrusted to the Hopenet model described in section 2.3. However, during tests this model highlighted some difficulty to make constant and correct estimations in situations in which the head was too much turned (an estimation of these limits can be similar to the ones of the BIWI dataset [98]: $\pm 75^{\circ}$yaw, $\pm 60^{\circ}$pitch, $\pm 50^{\circ}$roll). Therefore, it was required to impose some limitation in its use exploiting only information extracted until the head orientation estimation. As a result, four checks based solely on the head's key points have been developed. By doing so, not only was it possible to avoid employing the model in problematic circumstances, but it also made possible to extract some useful data without directly estimating the orientation. The checks define if portions of the head related with the face are visible enough to proceed with the orientation estimation. The checks are reported in algorithm 3456:

Figure 3.9: Yaw Pitch and Roll rotations



Algorithm 3: Head Horizontal Visibility

if $(conf[eye\_l] \neq$ "$removed$" ***and*** $conf[ear\_r] \neq$ "$removed$"$)$ ***or*** $(conf[eye\_r] \neq$ "$removed$" ***and*** $conf[ear\_l] \neq$ "$removed$"$)$ **then**
  return $True$
end
return $False$

Algorithm 4: Head Vertical Visibility

if $conf[nose] \neq$ "$removed$" ***and*** $conf[head] \neq$ "$removed$" **then**
  return $True$
end
return $False$

Algorithm 5: Is Head Front

if $((conf[ear\_l] \neq$ "$removed$" ***and*** $conf[ear\_r] \neq$ "$removed$"$)$ ***and*** $(kp[ear\_l][x] > kp[ear\_r][x]))$ ***or*** $((conf[eye\_l] \neq$ "$removed$" ***and*** $conf[eye\_r] \neq$ "$removed$"$)$ ***and*** $(kp[eye\_l][x] > kp[eye\_r][x]))$ **then**
  return $True$
end
return $False$

Algorithm 6: Face Keypoints Coherency

$eye\_l\_to\_head = ((kp[eye\_l][x] - kp[head][x])^2 + (kp[eye\_l][y] - kp[head][y])^2)$
$eye\_r\_to\_head = ((kp[eye\_r][x] - kp[head][x])^2 + (kp[eye\_r][y] - kp[head][y])^2)$
$eye\_l\_to\_eye\_r = ((kp[eye\_l][x] - kp[eye\_r][x])^2 + (kp[eye\_l][y] - kp[eye\_r][y])^2)$
if $eye\_l\_to\_head < eye\_r\_to\_head$ **then**
  $ref\_value = eye\_l\_to\_head/value$
else
  $eye\_r\_to\_head/value$
end
if $eye\_l\_to\_eye\_r < ref\_value$ **then**
  return $False$
end
return $True$

The first two are very coarse controls that only ensure that specific vertical and horizontal portions of the face are visible by checking that couple of keypoints have not been flagged as "removed". If a person's keypoints do not pass those checks, the head orientation estimation will not be done on that person and its result will contain the value used to indicate "removed". The third defines if the head is facing or is showing a back related part of it to the camera. The fourth was needed because Alphapose, as other body pose estimators, tries to define the position of keypoints that are not directly visible, therefore, it can make the other checks unreliable. This happens when the head yaw is near $\pm 90^{\circ}$. In this particular positions, eyes' keypoints will be very close to each other, for this reason, it can be used as a clue of the presence of this limit case. To avoid that this close distance derives from the dimension of the head (e.g., a far person with respect to the robot), instead of the presence of those particular poses, a fraction of that distance is compared with the greater one between left eye-head and right eye-head. The head orientation estimation will not be performed on a person whose keypoints fail the last two tests, and the result will include a value used to indicate "back".

The model used to estimate head orientation needs as input, images with only a single head. For this reason, starting from the input image, the latter are found by using the head bounding boxes provided by the preprocess module. In the end the cropped images are given to the model and the yaw, pitch and roll for each of them are retrieved. The response will contain both results of heads processed by the model and the ones blocked by checks.

3.3.1.4 Body Orientation

As reported in 1.7.1, body orientation can express plentiful nonverbal information about engagement whether it is about initiating or maintaining an interaction. Also, as delineated in 1.9.1 it can be a fundamental cue in collaborative interactions. For these reasons, body orientation has been considered in this project.

Body orientation block has the role to estimate the angle of people's body with respect to the robot. To do so, it requires the input image, the keypoints and the bounding boxes

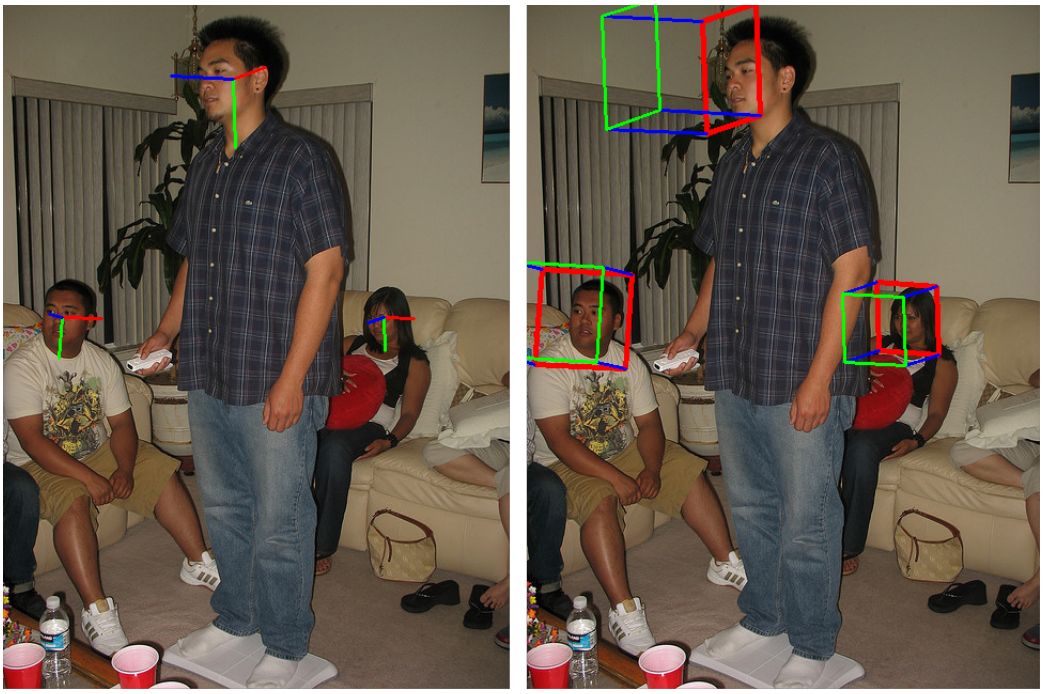


Figure 3.10: Visual plotting of Head Orientation module results. (a) heads reference frame rotated by the estimated angles (x pointing to right, y to the ground, z out of the screen) (b) 3D cubes to enhance the perception of the angles estimated. Yaw: green, Pitch: red, Roll: blue

extracted by the preprocess component.

The estimation of this orientation is entrusted to the MEBOW model described in section 2.2.1. The orientation estimated is considered as defined by the authors when taken as single component result, while its value is redefined during the aggregation. The orientation used in MEBOW is the angle between the projection vector of the chest facing direction (C) onto the y-z plane and the direction of the axis z, where the x, y, z vectors are defined by the image plane and the orientation of the camera. Given a 3-D human pose, the chest facing direction C can be computed by $C = T \times S$, where S is the shoulder direction defined by the vector from the right shoulder to the left one, and T is the torso direction defined by the vector from the midpoint of the left- and right-shoulder joints to the midpoint of the left- and right-hip joints. MEBOW provides as result the most probable along 72 bins. The distribution of those is pictured in Figure 2.7. During tests, this model revealed some estimation issues when working with bodies observed from behind (e.g., from 270°to 0°and from 0°to 90°), returning completely incorrect results in some instances. On the other hand, when bodies were front-facing the camera, the results appeared to be consistently accurate and very fine-grained. For this reason, as done for the head orientation module, some checks have been used to avoid using the model in problematic situation while still extracting information that may be useful for the final aggregation. The checks define if portions of the body are visible enough to proceed with the orientation estimation. The checks are reported in algorithm 789:

Algorithm 7: Upper Body Vertical Visibility

if $(conf[s\_l] \neq$ "$removed$" ***and*** $conf[hip\_l] \neq$ "$removed$") ***or*** $(conf[s\_r] \neq$ "$removed$" ***and*** $conf[hip\_r] \neq$ "$removed$") **then**
  return $True$
end
return $False$

Algorithm 8: Is Body Front

if $((conf[s\_l] \neq$ "$removed$" ***and*** $conf[s\_r] \neq$ "$removed$") ***and*** $(kp[s\_l][x] > kp[s\_r][x]))$ ***or*** $((conf[hip\_l] \neq$ "$removed$" ***and*** $conf[hip\_r] \neq$ "$removed$") ***and*** $(kp[hip\_l][x] > kp[hip\_r][x]))$ **then**
  return $True$
end
return $False$

Algorithm 9: Are Shoulder Keypoints Superimposed

if $(conf[s\_l] \neq$ "$removed$" ***and*** $conf[s\_r] \neq$ "$removed$") ***and*** $(kp[s\_l][x] == kp[s\_r][x]$ ***and*** $kp[s\_l][y] == kp[s\_r][y])$ **then**
  return $True$
end
return $False$

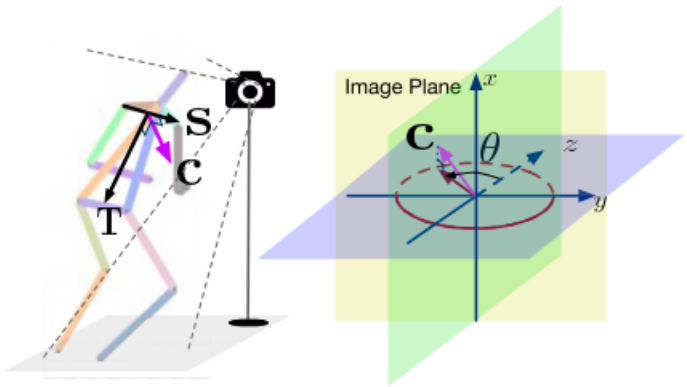


Figure 3.11: Body orientation definition used in MEBOW. Image taken from the paper [313]

The first ensure that the left or right side of the torso frontal view is visible to the camera. The body orientation estimation will not be performed on that person and the result will contain the value used to indicate "removed" if the keypoints do not pass that check. The second specifies whether the camera is viewing the body from the front or back. The third step determines whether the left and right shoulder keypoints are completely overlapping. This check had to be done because Alphapose can estimate keypoints of parts even when they are not directly visible. However, when this happen for shoulders and hips, very often is one of them that occlude its counterpart, therefore, the keypoint estimated but not directly seen is superimposed to the one of the opposite side. In these situations, the body orientation is very close to the limits imposed before but will pass the first two checks easily. If a person's body do not pass the latter check, his body orientation estimation will

not be done and his result will contain the value used to indicate “back”.

The model used for body orientation estimation, needs as input, images with a single person. For this reason, starting from the input image, they are obtained by cropping it using the full body bounding boxes found from the preprocess module, they are then resized so to be compatible with the model, and in the end passed to the model itself. The result is multiplied by 5 to obtain the information in degrees.

The results of each person, whether deriving from the model or from the controls, are ultimately sent back as response.

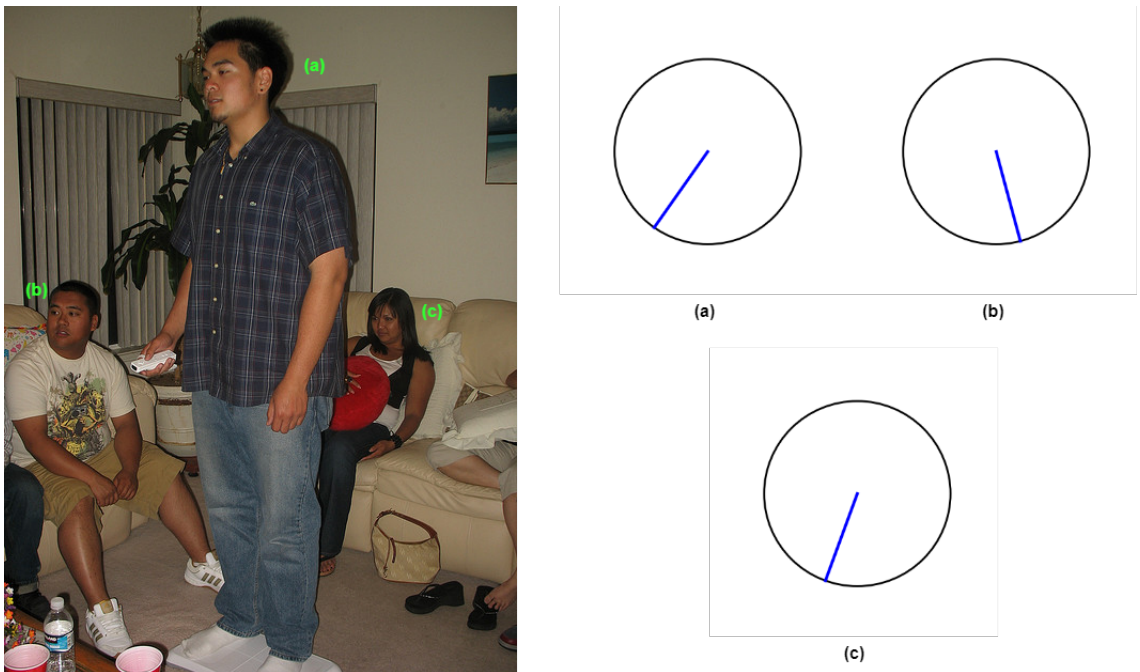


Figure 3.12: Body Orientation results. (a) 215°. (b) 165°. (c) 200°

3.3.1.5 Pose Classification

Nonverbal cues derived from a person’s posture, as stated in 1.7.1, can reveal a great deal about a person’s feelings and attitudes. Additionally, they can tell much about social relationships and the structure of social interaction. From all the possible information interpretable, the body openness is directly connected to engagement. In fact, this particular nonverbal cue can unconsciously influence the perception and will of a person, making it more or less prone to start, continue or even end a conversation with the observed person. Following 1.7.1 and [229], different body parts can express this cue, among those the upper body and hands have been considered in this project.

A specific dataset for each type of classification proposed did not existed. For this reason, custom datasets have been built putting together others found online and by annotating multipurpose datasets as COCO [193] and Open Images [183][34][236]. However,

because there was not an infinite amount of time available for the realization of those four, it was possible to create custom datasets that only contained sufficient data to train decent models. Having few data, it was decided to use pre-existing models and fine-tuning them on the custom datasets created. To choose which one to utilize, their performance on ImageNet-1K have been analyzed in relation with the number of parameters used. The models that were taken into consideration had pre-trained weights that could be found in Pytorch hub. In the end ConvNext [199] tiny was chosen.

Training was done using Pytorch, for this reason it was possible to use its data transforms to implement data augmentation. The data transform composed for training randomly change the brightness, contrast and saturation of the image (ColorJitter with 0.3 to each cited parameters), then it crops a random portion of that and resize it to a given size (RandomResizedCrop, scale is the RCS low hyperparameter studied), then, with a given probability (0.5) it does a horizontal flip and in the end, it normalizes the image. The one used for validation instead only resize, crops at the center so to have the correct dimensions and then it normalizes the input image.

Additionally, because balancing the datasets' classes would have excluded some examples, it was utilized the option to specify weights for each class. By doing so examples of a class with greater weight will be considered more valuable to the model. The weights were computed as:

$$weight\_class_n = 1 - \frac{num\_examples\_class_n}{num\_tot\_examples} \tag{3.2}$$

For every classification type, a hyperparameter study has been conducted. These studies have been done using a library called optuna [6] and exploiting its Tree-structured Parzen Estimator (TPE) algorithm [37][36] for sampling the hyperparameter to use in each trial. That sampler is based on independent sampling. On each trial, for each parameter, TPE fits one Gaussian Mixture Model (GMM) $l(x)$ to the set of parameter values associated with the best objective values, and another GMM $g(x)$ to the remaining parameter values. It chooses the parameter value x that maximizes the ratio $l(x)/g(x)$.

The first study conducted pointed to minimize the validation loss across the trials, but because the model tried overfitted and diverged very rapidly with that dataset, the results obtained seemed not to improve the starting situation. Then it was tried to maximize the accuracy across trials obtaining some improvements, therefore, it was decided to continue in this way also for the remaining models. The initial study consisted of 150 trials, each of which utilized early stopping and consisted of no more than 100 epochs of training. In the subsequent studies, fewer trials were performed, the searching space was reduced, and the training settings were altered due to the fact that a study of the previous kind took more than 72 hours to complete, and the machine used for this task was available only during non-working hours. The other studies consisted of 100 trials with trainings of no more

than 50 epochs and early stopping. For each trial, the loss function was cross-entropy. The hyperparameters composing the searching space can be seen in Tables 3.13.2.

	Values
Learning Rate	[1e-5, 1e-1]
Optimizer	Adam, SGD
Batch_size	16, 32, 64
Rc Scale Lower	0.5, 0.6, 0.7, 0.8, 0.9

Table 3.1: Hyperparameters composing the searching space for the study body upstanding, sitting, lying

	Values
Learning Rate	[1e-5, 1e-1]
Optimizer	Adam, SGD
Batch_size	32, 64
Rc Scale Lower	0.08, 0.5, 0.9

Table 3.2: Hyperparameters composing the searching space for the studies of body open, closed; hands open, closed; hands palm, back

The hyperparameters found were then used in the final training of their respectively models. Each training was composed by a maximum of 150 epochs and a patience of 50 epochs.

Body classification: upstanding, sitting, lying

For this classifier, different datasets have been merged together. IASLAB-RGBD Fallen Person Static Dataset [17] contains 360 images with lying people. Freiburg Sitting people dataset [224] constitutes a dataset with 200 images of six different people sitting in multiple viewpoints and in a wide range of orientations. E-FPDS dataset [185] consists of 6982 images, with a total of 5023 falls and 2275 non falls corresponding to people in conventional situations (standing up, sitting, lying on the sofa or bed, walking, etc); from it only the one containing fallen people have been considered. COCO dataset [193] is a large-scale object detection, segmentation, and captioning dataset with 123287 images and 886284 objects instances. From it, 165834 images have been extracted by cropping people using their bounding boxes provided in the annotation of the dataset. The images were then processed by the body orientation estimation module, and because only people who were facing the camera were of interest, only those with results between 90°and 270°were taken into consideration. Images with width · height less than 2000 were also not taken into consideration because it was very difficult to understand them. In the end, their number was reduced to 62535. COCO provide the dataset in already divided training and validation splits. The previous process was done on each split without merging them. After that, I annotated 10246 images for the training split, classifying 3614 as standing, 1322 as sitting, 100 as lying, and 5210 as unusable. The latter class refers to images for which I preferred not to assign a class because it was either impossible to do so or the image was very ambiguous. For the validation split instead, 1747 images have been annotated by me from which 483 as upstanding, 262 as sitting, 6 as lying, 996 as unusable.

In order to have nearly an 80%/20% division between training and validation split, the datasets were merged as follows:

Train:

Upstanding: COCO annotated (3614) | TOT = 3614

Sitting: COCO annotated (1322) + Freiburg test (130) | TOT = 1452

Lying: COCO annotated (100) + E-FPDS train (3822) | TOT = 3922

Validation:

Upstanding: COCO annotated (483) | TOT = 483

Sitting: COCO annotated (262) + Freiburg train (70) | TOT = 332

Lying: COCO annotated (6) + E-FPDS val (762) + IASLAB (360) | TOT = 1128

Results of the hyperparameter study done for this classifier can be found in Figure 3.13 and Table 4.19 with the set used for the final training highlighted. The results of the training can be found in Figure 3.14.

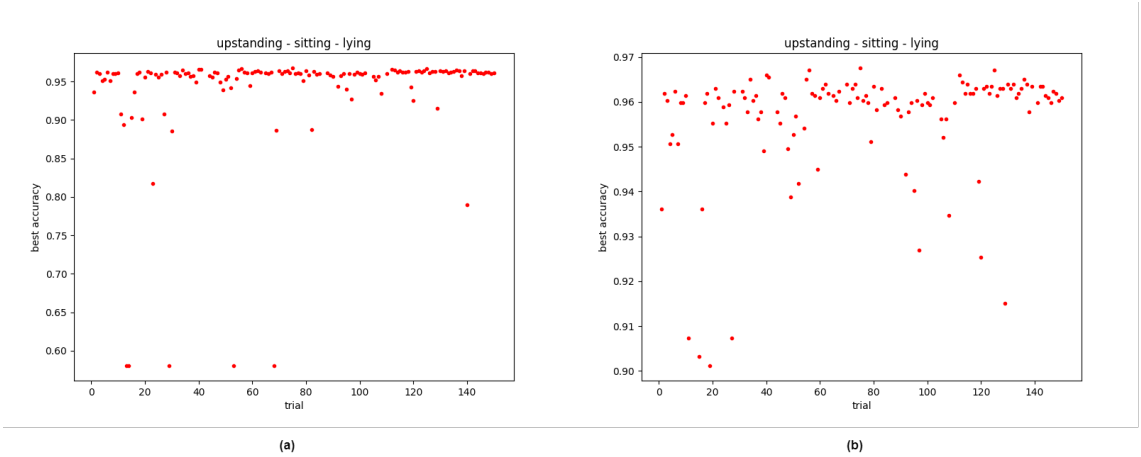


Figure 3.13: Upstanding-sitting-lying hyperparameter study results. (a) original plot. (b) plot limited to value > 0.9 ta have a more refined view

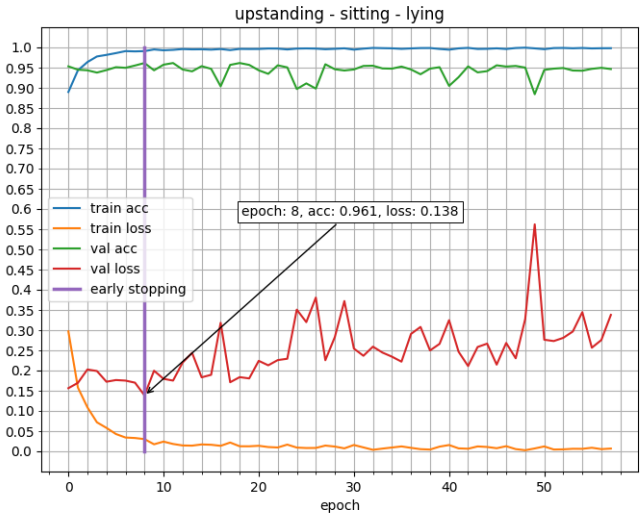


Figure 3.14: Upstanding-sitting-lying training results

Body classification: open, closed

For this classifier, only the COCO dataset has been used. As done previously, from images of this dataset single person have been cropped, then only picture with person oriented between 90°to 270°(using body orientation module) and having height · width > 2000 have been considered. Those have been annotated by me using clues derived from Chapter 1 and [229] that may indicate an open or closed body pose. In particular my decision was driven mostly by observing the areas related with the torso and head, observing if they were occluded in some way by hands and arms. In the end 11476 images were annotated for the training split from which 2988 were classified as open, 3206 as closed 5282 as unusable, where the latter class refers to images for which I preferred not to assign a class because it was either impossible to do so or the image was very ambiguous. For the validation split instead, 2662 images have been annotated from which 670 were classified as open, 821 as closed and 1171 as unusable.

Figure 3.15: Example of images taken from the various datasets. (a) COCO. (b) IASLAB. (c) E-FPDS. (d) Freiburg

Results of the hyperparameter study done for this classifier can be found in Figure 3.16 and Table 4.20 with the set used for the final training highlighted. The results of the training can be found in Figure 3.17.

Hands classification: palm, back

For this and the next classifiers, four datasets have been used. The NUS hand posture datasets I [238] consists of 10 hand posture classes, 24 sample images per class, which are captured by varying the position and size of the hand within the image frame. The NUS hand posture datasets II [239] is a 10 class hand posture dataset in which the postures are shot in and around National University of Singapore (NUS), against complex natural backgrounds, with various hand shapes and sizes; the postures are performed by 40 subjects, with different ethnicities, both males and females and in the age range of 22 to 56 years. MNIST Sign language [300] follows the structure of other MNIST dataset but

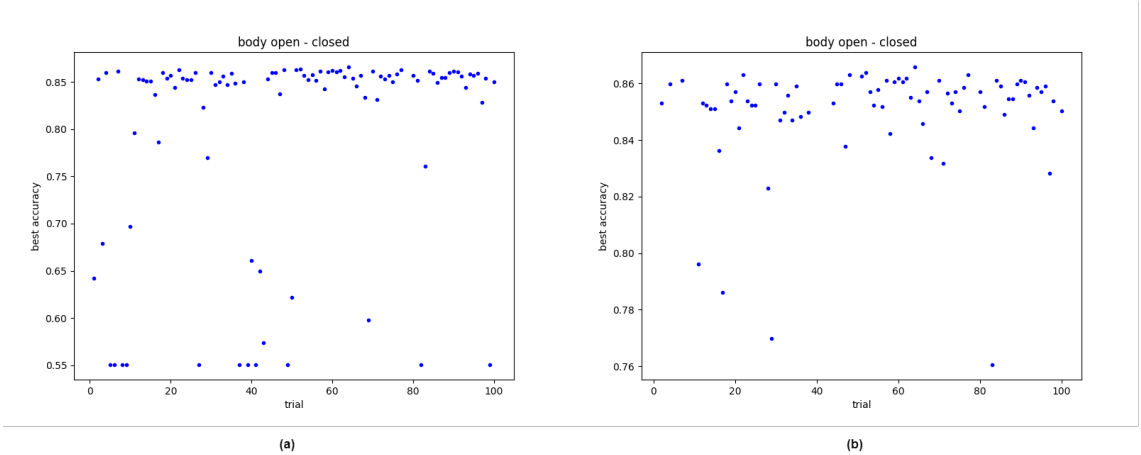


Figure 3.16: Body open-closed hyperparameter study results. (a) original plot. (b) plot limited to value > 0.75 ta have a more refined view

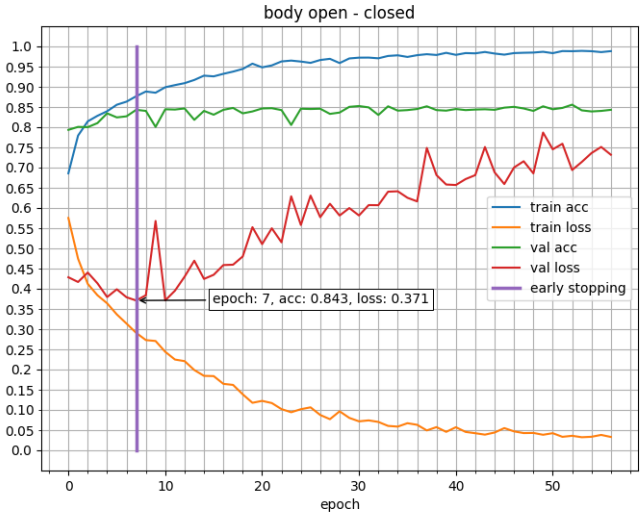


Figure 3.17: Body open-closed training results

focus on the American sign language letter (24 classes of letters excluding J and Z which require motion). Open Images is a dataset of 9 million images annotated with image-level labels, object bounding boxes, object segmentation masks, visual relationships, and localized narratives; it contains a total of 16 million bounding boxes for 600 object classes on 1.9 million images.

Because the first three datasets include clearly defined signs made with hands, it was possible to map those to a particular class. In this case the mappings can be found in Table 3.3 for NUS I, II and MNIST signs. MNIST sign language dataset was made by cropping and augmenting another dataset that can be found here [210]. For this reason, MNIST sign language have not be considered in the composition of the validation dataset.

From MNIST sign, 27455 training images could be used, but considering that for each image of the starting dataset about 50 have been created using data augmentation, only

a portion of them have been considered in the end. An algorithm has been developed to avoid the situation in which only poses belonging to the same class would have been taken into account by sampling a portion of the data. This, first finds out the number of images belonging to the new class with the smaller number of elements, then it computes the ratio between that number and the one of each classes. This ratio will be multiplied to the number of elements of each type of the original MNIST, finding out the number of elements that will be taken when randomly sampling each of those types. Each image sampled will be take part of the class defined following the map.

After mapping and sampling there were 2734 palm and 2103 back and 2733 neutral. The latter was a third class used in an initial approach to the problem, later it was decided to unify neutral and palm due to their similarity and because early training tests showed that the problem could be too complex for the amount of data.

From NUS I and II, after following the map 3.3, it was possible to have 1682 images classified as palm, 0 as back for the training split, 558 palm and 0 back for the validation.

From Open Images instead, example have been retrieved by cropping hands following the provided bounding box annotations, from which only the ones with width $\cdot$ height > 2000 were considered. Those have later been classified by me obtaining in the end 717 palm and 1283 back for training, 630 palm and 1451 back for validation (having the limitation to use MNIST only for training it was given priority to the validation set).

Train:

Palm: Open Images (717) + MNIST s (5467) + NUS I,II train (1682) | TOT = 7866

Back: Open Images (1283) + MNIST s (2103) + NUS I,II train (0) | TOT = 3386

Validation:

Palm: Open Images (630) + NUS I,II val (558) – TOT = 1188

Back: Open Images (1451) + NUS I,II val (0) – TOT = 1451

Results of the hyperparameter study done for this classifier can be found in Figure 3.18 and Table 4.21 with the set used for the final training highlighted. The results of the training can be found in Figure 3.19.

Hands class: open, closed

For this classifier, the same datasets as 3.3.1.5 have been used, but defining different maps than can be found in Table 3.3. From NUS I & II, 372 image were mapped as open and 1309 as closed for the training split, 124 open and 435 closed for validation. From MNIST 2214 images were classified as open, 2878 as closed and 2879 as partially closed. The latter was a third class used in an initial approach to the problem, later it was decided to unify closed and partially closed due to their similarity and because early training tests showed that the problem could be too complex for the amount of data.

Algorithm 10: Compute element to sample for each class

```
num_open, num_closed, num_palm, num_back = 0
foreach old_class ∈ old_classes do
    if map_OC[old_class] == "open" then
        num_open += num_ele(old_class)
    else
        num_closed += num_ele(old_class)
    end
    if map_PB[old_class] == "palm" then
        num_palm += num_ele(old_class)
    else
        num_back += num_ele(old_class)
    end
end
num_min_OC = min(num_open, num_closed)
num_min_PB = min(num_palm, num_back)
O_ratio = num_min_OC/num_open
C_ratio = num_min_OC/num_closed
P_ratio = num_min_PB/num_palm
B_ratio = num_min_PB/num_back
foreach old_class ∈ old_classes do
    num_old_class_train_OC = 0
    num_old_class_train_PB = 0
    if map_OC[old_class] == "open" then
        num_old_class_train_OC = num_old_class · 0_ratio
    else
        num_old_class_train_OC = num_old_class · C_ratio
    end
    if map_PB[old_class] == "palm" then
        num_old_class_train_PB = num_old_class · P_ratio
    else
        num_old_class_train_PB = num_old_class · B_ratio
    end
    element_list = copy(list_file_in_dir(images_class_n_path))
    for j ← 0 to num_old_class_train_OC do
        random_element = random_choice(element_list)
        copy_to_dir(random_element, train_new_class_m_path)
        remove(random_element)
    end
    for j ← 0 to num_old_class_train_PB do
        random_element = random_choice(element_list)
        copy_to_dir(random_element, train_new_class_m_path)
        remove(random_element)
    end
end
```

For what concerns Open Image dataset, the same image used in 3.3.1.5 have been classified by me obtaining 922 images classified as open and 876 as closed for the training split, 1173 as open and 959 as closed (having the limitation to use MNIST only for training it was given priority to the validation set).

Train:

Open: Open Images (922) + MNIST s (2214) + NUS I,II train (372) – TOT = 3508

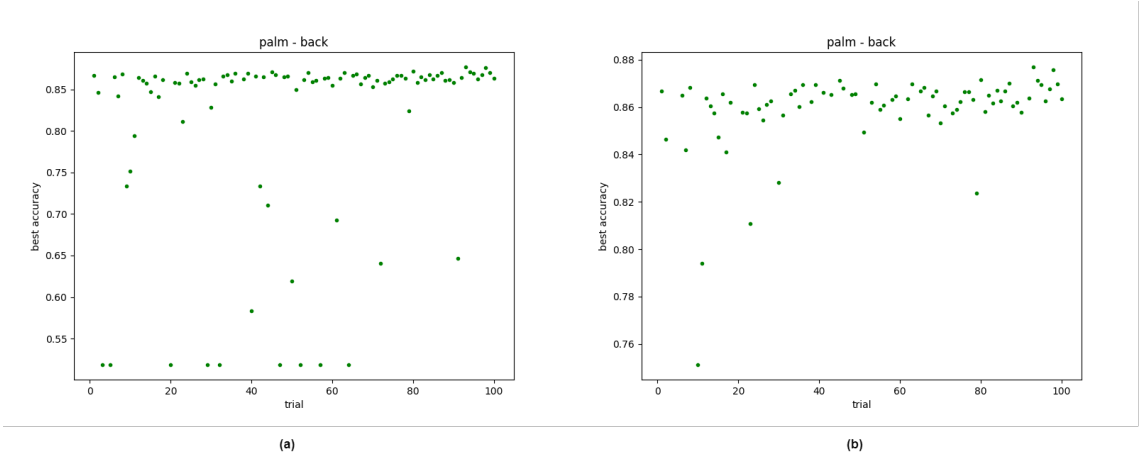


Figure 3.18: Hands palm-back hyperparameter study results. (a) original plot. (b) plot limited to value > 0.75 ta have a more refined view

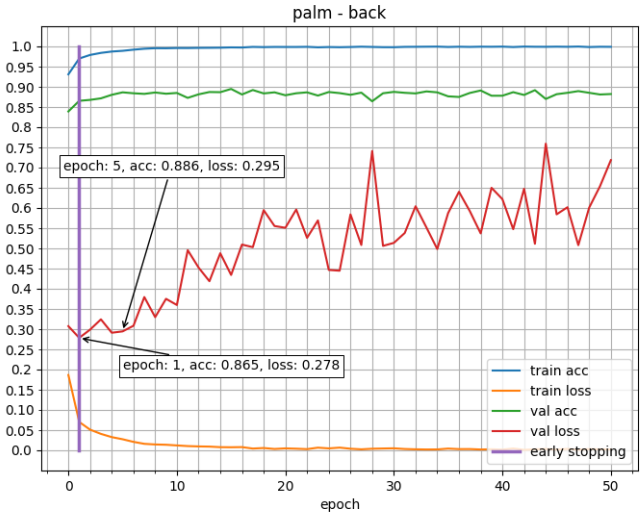


Figure 3.19: Hands palm-back training results

Closed: Open Images (876) + MNIST s(5757) + NUS I,II train(1309) – TOT=7942

Validation:

Open: Open Images (1173) + NUS I,II train (124) – TOT = 1297

Closed: Open Images (959) + NUS I,II train (435) – TOT = 1394

Results of the hyperparameter study done for this classifier can be found in Figure3.22 and Table 4.22 with the set used for the final training highlighted. The results of the training can be found in Figure 3.22.

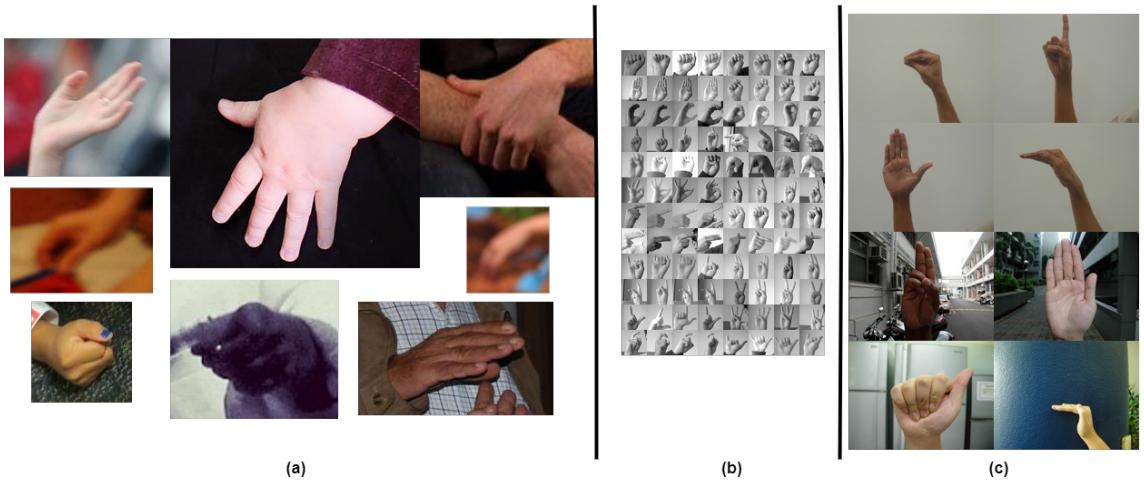


Figure 3.20: Example of images taken from the various datasets, their size has not been changed. (a) Open Images. (b) MNIST sign. (c) NUS I & II

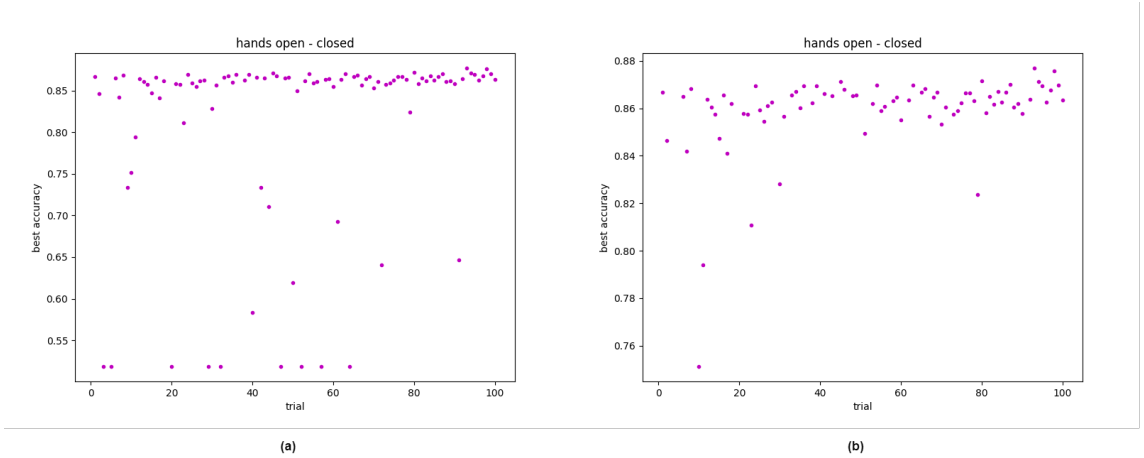


Figure 3.21: Hands open-closed hyperparameter study results. (a) original plot. (b) plot limited to value > 0.75 ta have a more refined view

3.3.1.6 Aggregation

The aggregation module has the role to combine the other component's results to obtain a value that indicates the Propensity For Interaction (PFI) of a person. The calculation of this element needs all previous results, for this reason, if a needed component has not been requested its result will still be calculated, however it will not be sent in the responses. The aggregation happens using three different groups of elements: components results values, components results confidences and aggregation weights. In the latter, three sets are defined respectively for values, confidences and head orientation. Every set sum to 1.

In the first step of the aggregation module, the components results values are normalized so to have a single value for each element and so that they are between 0 and 1. The head orientation results are first singularly normalized considering approximately the limit imposed in that module ($\pm 70°$yaw, $\pm 60°$pitch, $\pm 50°$roll) and then they are used to compute

Figure 3.22: Hands open-closed training results

NUS I		
Old class	Open-Closed	Palm-Back
G1	O	P
G2	O	P
G3	O	P
G4	P	P
G5	C	P
G6	P	P
G7	C	P
G8	P	N
G9	P	N
G10	O	N

NUS II		
Old class	Open-Closed	Palm-Back
a	C	P
b	O	P
c	P	P
d	C	P
e	P	N
f	O	N
g	O	N
h	C	P
i	P	N
j	P	P

MNIST sign		
Old class	Open-Closed	Palm-Back
0	C	P
1	O	P
2	P	D
3	D	D
4	C	P
5	O	D
6	C	B
7	C	B
8	C	P
10	C	D
11	C	P
12	C	P
13	C	P
14	P	D
15	C	N
16	C	N
17	C	D
18	C	P
19	C	D
20	C	P
21	C	P
22	P	P
23	C	N
24	C	P

Table 3.3: Mapping from old dataset classes to new ones used in this project. O:open, C:closed, P:partially-closed(only in op)/palm(only in pb), B:back, D:depends(used when in a category, there was some examples recognizable as one type while others as another one)

the final value as:

$$ho\_result\_processed = 1 - \sum ho\_results \cdot ho\_weights \tag{3.3}$$

Body orientation results are firstly modified so to be 0°when the person is completely facing the camera and so that by there is no difference in turning clockwise or counter-clockwise and then it is normalized considering that now the maximum value obtainable is 90°. The classification results instead are mapped into fixed values. It will be given a very low one to results that indicates the absence or impossibility of engagement such as

closed, back, lying and very high values to the ones that instead may indicates the presence of it such as open and palm. To the class sitting instead, a value that is nor good nor bad have been given based on the fact that when seated, the body is constrained and it is more difficult to make movements with it, therefore influencing the amount of nonverbal information expressible. This fact is taken in consideration also by adding to, and decreasing by, a fixed value respectively the head and body orientation value weights, due to the fact that the head is more likely to be less constrained than the body and, for this reason, it could be more expressive in that particular situation. When lying instead, a person's body is almost completely constrained and therefore a fixed low value is given to replace the head and body orientation ones. When a component's result has not been estimated (e.g., the body part in interest was not visible), was flagged as "removed" or its confidence is lower than the one specified in the aggregation module, the element is flagged as "not to use" by setting its confidence to a fixed value. Instead, a very low value is used as the component's replacement value when it has been flagged as "back", and at the same time a very high value is set for its confidence. This has been done on the idea that an important information is still being expressed but the meaning of that information is that the component indicates the absence of engagement.

After that, the value weights of elements flagged as "removed" are summed and re-distributed among all the remaining ones, giving more value to elements with a greater starting weight.

Algorithm 11: Aggregation module weight re-distribution

```
if (weight_lost > 0) then
    foreach weight ∈ components_values_weight do
        if weight ≠ not_to_use_value then
            weight = weight/(1 − weight_lost)
        end
    end
end
```

In the end, the PFI value is computed as the weighted sum of the elements' value not flagged as "not to use", and its confidence as the weighted sum of the elements' confidence not flagged as "not to use".

PFI value and its confidence have to be considered along while interpreting the final result because the first, having the re-distribution of the weights, define a value between 0 and 1 that define an amount of propensity for interaction given the visible or evaluable body parts, therefore it can reach great results even with only few computed elements, while the confidence does not have the re-distribution of the weights and therefore it loses the ones marked as "not to use" defining a value that gives information about the reliability of the aggregation, so it would be low when estimations confidences are low but also when only for a few components were possible to make the estimation.

Algorithm 12: Aggregation module PFI value and confidence computation

```
PFI_value = 0
PFI_confidence = 0
i = 0
while i < num_components do
    if components_weight[i] ≠ not_to_use_value then
        PFI_value += components_value[i] · components_value_weight[i]
        if components_weight[i] ≠ back_value then
            PFI_confidence +=
              components_confidence[i] · components_confidence_weight[i]
        else
            PFI_confidence += components_confidence_weight[i]
        end
    end
    i += 1
end
```

Chapter 4

Tests and Results

During tests, the system worked on a computer with CPU: i7-12700k, GPU: RTX 3080ti 12Gb, RAM: 32Gb (3200MHz). The images retrieved were captured by the stock camera of a temi V2 (13MP, 120deg FOV) and by a simple smartphone camera (sensor: Sony IMX 398, 16MP).

Images have been directly captured at a resolution of 640x480 or they were re-scaled before being given to the system.

In tables, for classifiers' results, only the first letter of that class has been reported. Always in them, a lot of abbreviations are used and they stand for: bo: Body Orientation, ho: Head Orientation, usl: Upstanding/Sitting/Lying, boc: Body Open/Closed, hocl: Hand Open/Closed Left, hpbl: Hand Palm/Back Left, hocr: Hand Open/Closed Right, hpbr: Hand Palm/Back Right and PFI: Propensity For Interaction. When an estimation/classification has not be done, its value and confidence will be None. In the table with more than one PFI entries, they refer to different people of the same image in order from the left to right.

4.1 Individual component detection

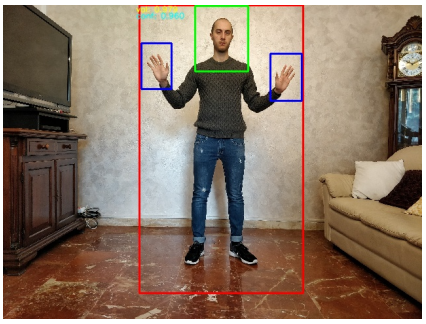
The final result of the implemented system depends only on the intermediate data extracted by the modules. This means that any problem related to one of those may have an impact on the correctness and reliability of the system itself. Therefore, the first test aims to show that each implemented component is able to work as defined. In each trial, one component is tested and the person detected is placed at a distance of 2.5-3m.

In Table 4.1(a) is depicted an image in which all components have been detected in their positive contribution value. In Tables 4.1(b), 4.4, and 4.5 classifiers module results have been reported while the one from orientation estimation can be found in Tables 4.2 and 4.3 (angles are expressed in degrees and the conventions used are the same reported in Chapter 3).

From those, it can be seen that at least in controlled situations like the ones reported, ev-

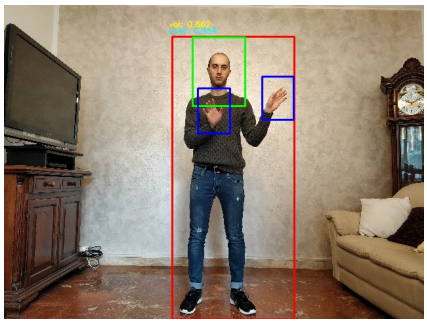
ery classifier can infer correctly between its own possible classes and that each orientation estimator can give acceptable results.

It can be noted that, when one classifier component is detected as having a non-positive value, the PFI decreases with respect to the one obtained in Table 4.1(a), and that the more the estimated angle from one of the orientation estimation modules is, the smaller the PFI will be.



(a)

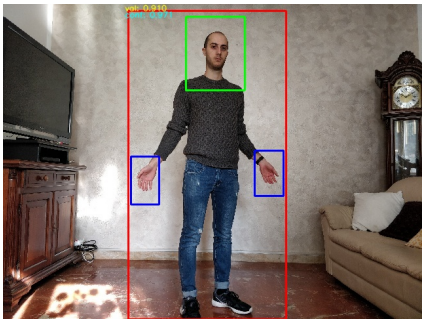
	Value	Conf
bo	180	0.914
ho	y:-5.288 p:-0.755 r:5.749	0.981
usl	u	0.999
boc	o	0.999
hocl	o	0.999
hpbl	p	0.636
hocr	o	0.999
hpbr	p	0.991
PFI	0.979	0.96



(b)

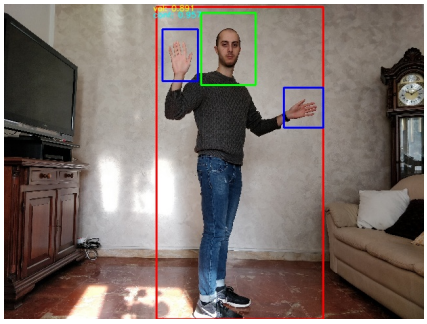
	Value	Conf
boc	c	0.99
PFI	0.862	0.954

Table 4.1: (a) Results with full positive contribution. (b) Body closed posture example



(a)

	Value	Conf
bo	190	0.909
PFI	0.91	0.971



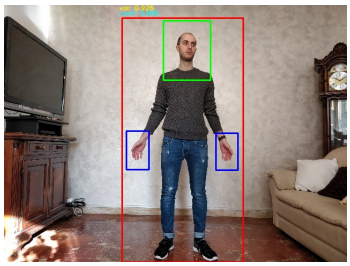
(b)

	Value	Conf
bo	200	0.889
PFI	0.891	0.957

Table 4.2: Body orientation examples

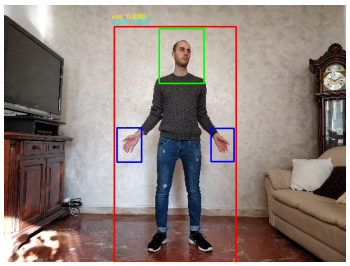
4.2 Generic global test

In this test, two sequences of images extracted from a video recorded by the robot have been labeled to verify the system's accuracy. The video was shot when we were testing the system and the chosen sequences depict respectively a situation where I was trying out the single modules and one in which up to 4 people were wandering around the room. The



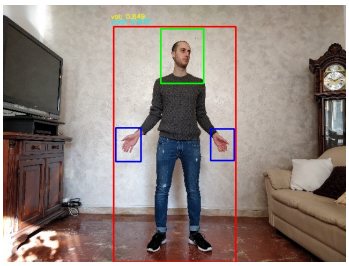
(a)

	Value	Conf
ho	y:-16.757 p:10.148 r:6.948	0.933
PFI	0.926	0.969



(b)

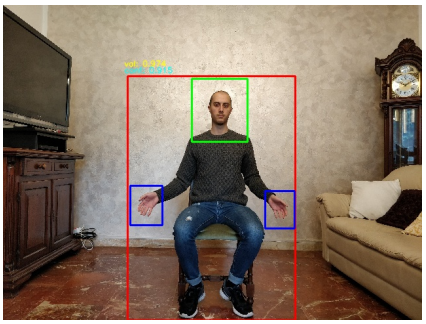
	Value	Conf
ho	y:-28.788 p:11.082 r:5.283	0.938
PFI	0.88	0.967



(c)

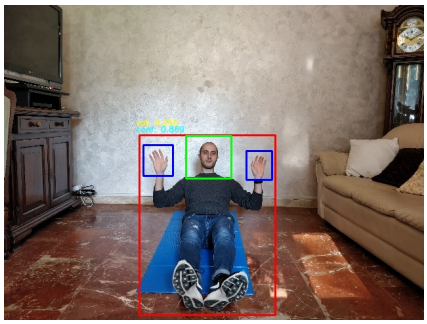
	Value	Conf
ho	y:-37.228 p:10.43 r:7.662	0.936
PFI	0.849	0.97

Table 4.3: Head orientation examples



(a)

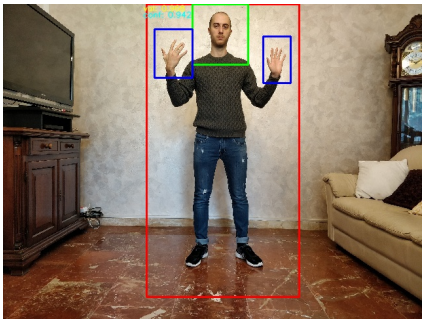
	Value	Conf
usl	s	0.909
PFI	0.974	0.915



(b)

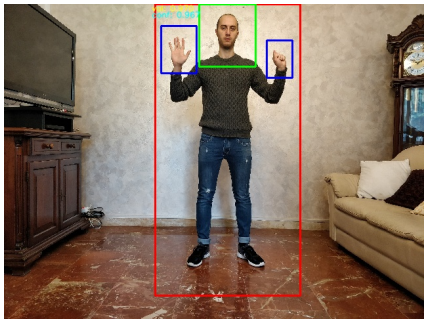
	Value	Conf
usl	l	0.889
PFI	0.255	0.869

Table 4.4: Sitting and lying down examples



(a)

	Value	Conf
hpbr	b	0.821
PFI	0.968	0.942



(b)

	Value	Conf
hocl	c	0.999
PFI	0.97	0.967

Table 4.5: (a) Example of hand back. (b) Example of hand closed

first case was chosen because it represents the optimal scenario but also contains situations in which I was discussing or interacting with colleagues out of the robot's field of view. The second sequence instead, was chosen because it represents a generic scenario in which people do not remain still and therefore their pose could be more natural and particular situations as person-to-person occlusion may occur. The frames labeled are 652 (334+318) and were taken every 0.5 seconds so to record more than 2 and a half minutes per sequence, but because one label may be given for each person in each frame, the number of labels

for a single component may be greater than the number of frames.

Because I am the one that labeled all the images and I am not an expert in the nonverbal communication field, the test has the objective of providing data to highlight the strengths or weaknesses of the system components more than showing that the final results are good. However even the PFI has been labeled and therefore its accuracy is provided, but, that data has to be considered with in mind the assumption that it may be inflated by a bias derived from the fact that I built the system and for this reason, the reasoning used during the PFI labeling procedure could be similar to the basic idea of the system itself. In the labeling procedure, I attributed discrete values for each body part of interest. For the ones related to classifier modules, it was simply used "0" to sign the positive class and "1" for the negative (0,1,2 for usl). For the orientations and PFI modules, discrete maps have been defined considering the value obtained during some tests. The label "4" has been used to indicate the impossibility of labeling that part because it was occluded or too blurred or even when it was too ambiguous to decide with enough certainty. All the labels' maps have been reported in Table 4.6(a).

	Component total labels	Labels - results matches	Accuracy (%)
bo	873	726	83.16
ho	817	609	74.54
usl	1030	988	95.92
boc	680	624	91.76
hocl	526	332	63.12
hocr	544	368	67.65
hpbl	567	191	33.69
hpbr	648	251	38.73
PFI	855	769	89.94

(a)

	Label: 0	Label: 1	Label: 2
bo	[165°, 195°]	[120°, 160°) or (195°, 240°]	val <120°or val >240°or val = None
ho	[-20°, 20°]	[-70°, -20°) or (20°, 70°]	val <-70°or val >70°or val = None
usl	u	s	l
boc	o	c	-
hocl	o	c	-
hocr	o	c	-
hpbl	p	b	-
hpbr	p	b	-
PFI	val $\geq$ 0.7 and conf $\geq$ 0.7	val < 0.7 and conf < 0.7	-

(b)

Table 4.6: (a) Generic global test label maps. (b) Results generic global test

The result of the test can be found in Table 4.6(b). All components but the ones related to hands have performed well, with the boc and usl classifiers performing equal or even better than what obtained on the validation set (usl_val 96.1% vs usl_test 95.92% and boc_val 84.3% vs boc_test 91.76%). The results obtained by the orientation modules seem good having respectively around 83% and 74% accuracy for body and head, but they may have been influenced both positively and negatively by the coarse re-mapping used. The hands' classifiers instead, have performed poorly giving the opportunity to discuss what could be wrong with them. Some elements may lead us to think that the problem could be the dataset used. Firstly, the results obtained from the training procedure highlight a divergent behavior on the validation set that could be caused by the model's inability to generalize using the training data provided to it. Secondly, there is a not negligible difference in the results when the same model is applied to the right or left hand (about 5%), which may be caused by the unbalanced presence of examples of one of those types (even

if it was actually used the RandomHorizontalFlip to try to avoid this problem). Thirdly, the results of the pb classifier are so poor (left: 33.69% right: 38.73%) that they not only may indicate the model's inability to generalize from the training data but also that actually the model may have learned to distinguish two completely different things with respect to the palm-back categories. Additionally, the results may have been influenced by the labeling process used for this test. In fact, it was not easy to understand a small body part as a hand in the images provided, because their resolution after the undistortion process was only 379x329, therefore some labeling errors may have occurred.

Even if the hands' classifiers performed poorly, the PFI results have not been influenced so much because of the small importance given to them during the aggregation. Obtaining an accuracy of around 90%, in general, using the threshold proposed in Table 4.6(a) for the PFI, the system seems to be at least usable even in the current implementation.

4.3 System limitations

4.3.1 Body and head occlusions

The system is quite robust to occlusions and can manage well the absence of almost all body parts of interest. However, during tests, some particular behavior related to head or body occlusion have been found.

Regarding the head, the normal behavior of the system in situations in which it is not visible because, for example, it is out of the robot's field of view or is occluded by something, would be to consider it as something flagged as "removed". Instead, sometimes, due to the presence of even a little portion of the head or neck, the head is still detected, or at least the pose estimation modules infer the head keypoints even without being able to see it directly as shown in Table 4.7. However, this problem has been encountered only when the head is heavily occluded but a recognizable portion of it (even a very small one) is quite visible and may be solved by tuning the head keypoints threshold or by utilizing another pose estimation model.

While the head may figure as not detected, the body cannot, because the system will define as body the largest group of body parts of the same person. In Table 4.8 it can be noticed that the body orientation estimation is done even when a small portion of the torso is visible, and therefore, the keypoints have passed all visibility checks defined. Indeed, by plotting the keypoints estimated during the preprocess as depicted in Table 4.9, some body parts occluded by the chair are still present. However, in Table 4.9(b), while the keypoints around the heap have been plotted, their confidences were too low to be considered and for this reason, the bounding box defined during the preprocess does not include that portion. Always from Table 4.8, it can be seen that while a partial occlusion of the body allows safe estimations, an almost total obscure of it, can bring unexpected results. This situation mostly appears when a great portion of the body is occluded without the possibility to see

the extrema portion of itself and in those cases, the results will report a low PFI value or confidence. The great variation in PFI is given by the incorrect attribution of the class lying and by all the limitations imposed when that particular pose is detected. This erroneous prediction may be caused by the data used for the training, in fact, it may be the case that among all the examples, a great portion of the ones that depict only partially the body, they are labeled as lying. Additionally, it has to be considered that without depth information, when only the upper part of the body is visible, someone lying down in a vertical position it may appear similar to one standing, moreover when context information retrievable in the image is limited because of a tight bounding box. A way to avoid this may be to define another check exploiting body keypoints and their confidence, from the fact that the more a keypoint has been "guessed" because not directly visible, the smaller its confidence should be.

(a) (b)

Table 4.7: Examples of scenarios with occluded head

(a)

	Value	Conf
usl	l	0.717
PFI	0.255	0.866

(b)

	Value	Conf
usl	u	0.717
PFI	0.958	0.916

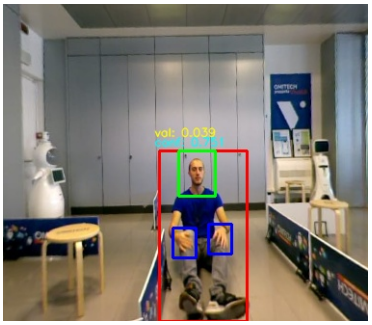
Table 4.8: Examples of scenarios with occluded body

4.3.2 PFI fluctuation

Another strange behavior encountered during tests is the fluctuation of the PFI value in consecutive frames as shown in Table 4.10. This peculiarity has been detected only when the person is in an ambiguous pose as in Table 4.10 or when he is at a distance of around

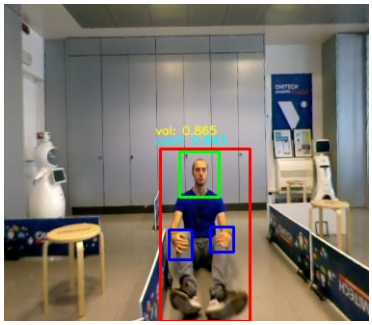
Table 4.9: Pose estimation on the same images of Table 4.8

4.5m and more. The first case has been analyzed previously, on the second instead, very few things can be said. In fact, at such a distance is quite understandable that at the current resolution, orientation modules can give coarser results and have less consistency, and that everything which regards hands may be less accurate due to their small dimension. To solve this problem, increasing the resolution may be an option but by doing so, the weight of each message will increase and may cause latency in requests and responses. Another way could be to vary the body parts threshold according to the dimension of the bounding boxes but if not done in the appropriate way, this may lead to the appearance of other strange behavior and unpredictable results in particular situations.



(a)

	Value	Confidence
PFI	0.039	0.751



(b)

	Value	Confidence
PFI	0.865	0.843

Table 4.10: Examples of a an ambiguous situation

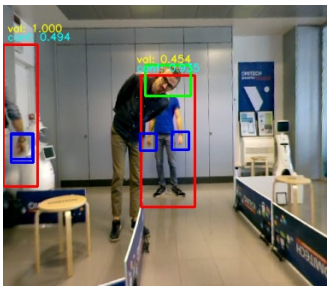
4.3.3 Person to person occlusion

A limitation inherited from the use of the body pose estimator exploited in the system is the inaccurate detection of people when they are overlapping. As it can be seen in Table 4.11, people are distinct as long as there is no clear overlap, but when the latter happens, one of them may not be detected or what will be given as result is the composition of the two body as one. However, for it to occurs, is needed an almost complete superimposition of the bodies or one of them has to have some body parts not visible and the other has to overlap the same parts on the first body. Because the situation is very specific, the problem does not occur often, and even when it may happen, most of the time it can work properly

as shown in Tables 4.12, but still, it has to be considered because the system may be used in a crowded situation. Unfortunately, because the problem derives from the body pose detection, the only way to avoid this problem is to use a different model for the same task.

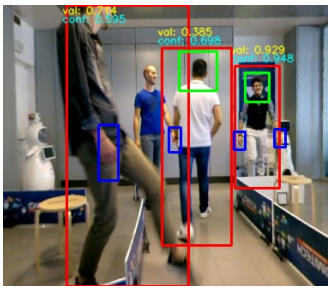
(a)

	Value	Confidence
PFI	0.836	0.728



(b)

	Value	Confidence
PFI	1	0.494
PFI	0.454	0.935



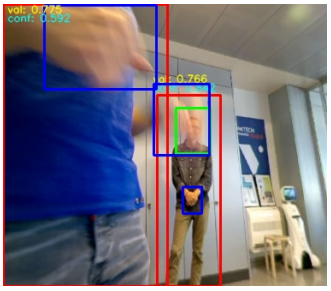
(c)

	Value	Confidence
PFI	0.714	0.595
PFI	0.385	0.698
PFI	0.929	0.948

Table 4.11: Examples in which problem related to person to person occlusion appeared

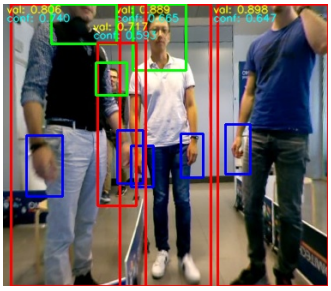
(a)

	Value	Confidence
PFI	0.522	0.429
PFI	0.398	0.798



(b)

	Value	Confidence
PFI	0.775	0.592
PFI	0.766	0.842



(c)

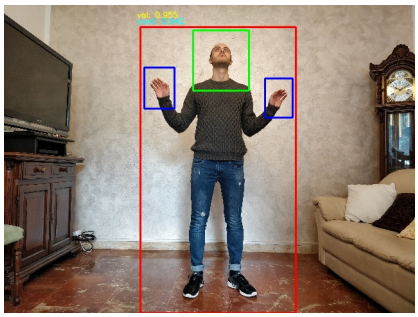
	Value	Confidence
PFI	0.806	0.74
PFI	0.717	0.593
PFI	0.889	0.665
PFI	0.898	0.647

Table 4.12: Examples in which problems related to person to person occlusion may have appeared but they did not

4.3.4 Currently Implementation

The system background idea derives from a simplification of what someone can analyze knowing some nonverbal communication concepts, but in the actual implementation, an important aspect is not exploited, or at least its implementation is still too coarse. Indeed, at this moment, the system does not consider time, it takes every single frame and the result obtained refers only to that. This may cause incorrect PFI value, or at least, results that may differ from what even a common person may have guessed. For example, in Table 4.13 the PFI value and confidence are high because each component (but the pitch) has been detected as having a positive contribution value at least in that precise frame. It may be the case that the person was ready for an interaction with the robot, but something happened, and his attention changed to something else, showing this by a great tilt of the

head. In particular situations, if this will bring the robot to start the interaction with the person, it may cause another change of focus in the person and if the initial situation was a dangerous one, the second may be even more because the robot may cause a distraction. Obviously, we are talking of a very specific situation that may appear only if the robot is inserted in specific contexts, but it highlights the concept. An improvement, in this case, may come from the introduction of something that considers intermediate results over a short amount of time.



	Value	Confidence
bo	180	0.903
ho	y:-4.395 p:25.51 r:3.271	0.924
usl	u	0.999
boc	o	0.999
hocl	o	0.941
hpbl	p	0.809
hocr	o	0.99
hpbr	p	0.786
PFI	0.955	0.942

Table 4.13: Example of a particular situation where even if almost all the component give positive contribution and therefore there is a high PFI, it would be the case to not interact with that person

4.4 Runtime Analysis

4.4.1 Individual Components

In this test, runtimes of individual components are reported. The trials aim to show the time spent by every single component, for this reason, the retrieved execution times are not relative to the exact module used in the final system, but only to those parts that actually compute their results, therefore a simplified version of the system has been used.

For this test 20 trials have been considered. Each of them used the simplified version of the system on an image. Each one of them has the same person repeated several times. The person used allows the detection of all his body parts. The runtimes information registered for each component and for each person detected were finally averaged. The results can be seen in Table 4.14 and a pie chart that highlights the execution time in percentage with respect to the total employed is shown in Table 4.15(a).

The situation is pretty clear. Preprocess and body orientation are the most time expensive components among all others, taking together more than 70% of the total registered. However, preprocess is done only one time at the start of the pipeline, therefore, when more people are detectable, its total runtime remains quite unaltered. An example of this can be seen in Table 4.15(b) and in the results of the next test.

As expected, because they exploit the same base model, classifiers have all similar execution time, however, having only the weights found during training to differentiate them

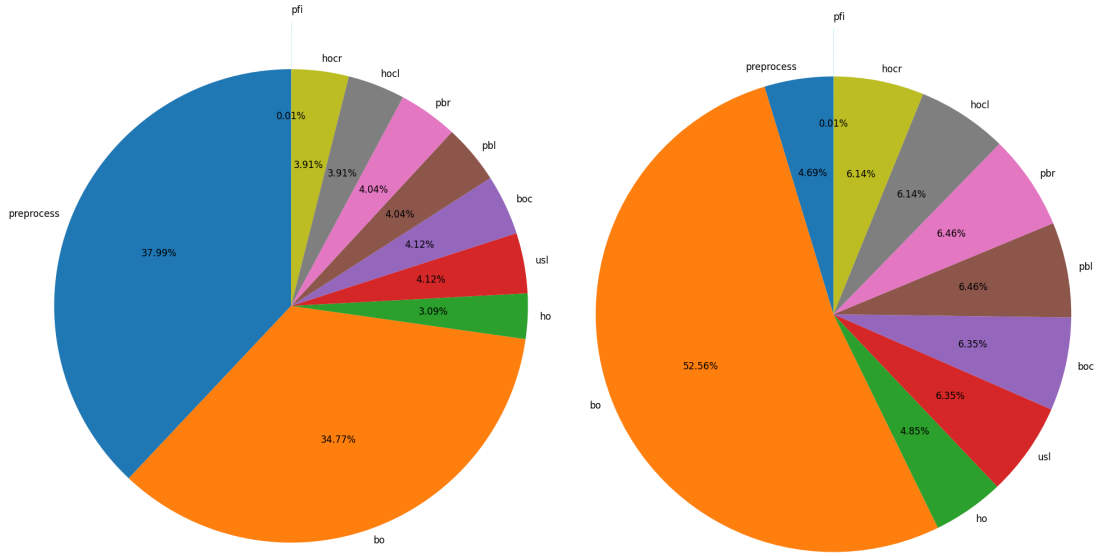
	Average time (s)
Preprocess	0.12205
Body orientation	0.11172
Head orientation	0.00994
Body usl	0.01323
Body oc	0.01323
Hand oc left	0.01299
Hand pb left	0.01299
Hand oc right	0.01256
Hand pb right	0.01257
PFI	0.00003

(a)

# people	Average time (s)
1	0.28818
2	0.45323
3	0.64819
4	0.86558
5	1.04030
6	1.27448
7	1.43309
8	1.66571
9	1.87075
10	2.04449
11	2.23349
12	2.45149
13	2.61075
14	2.80333
15	3.09728
16	3.19541
17	3.39053
18	3.67699
19	3.82651
20	3.98125

(b)

Table 4.14: (a) results of the Individual components test. (b) results of the Number of people detected test



(a) (b)

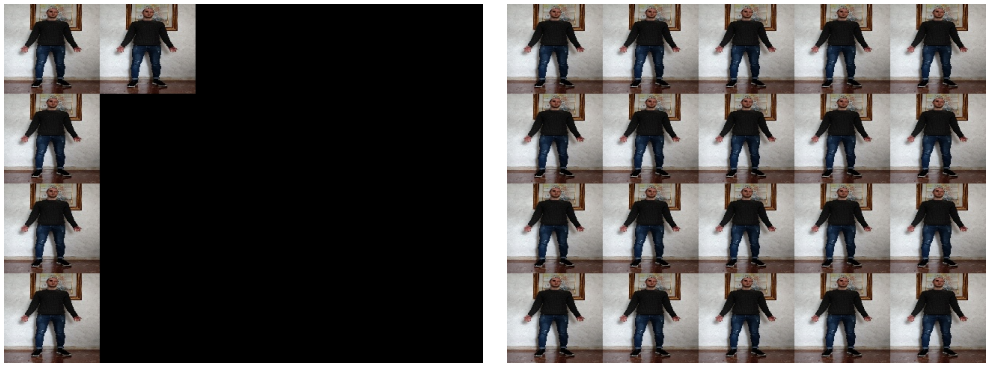
Table 4.15: (a) percentage of time spent by individual components compared to the total when working on only one person. (a) percentage of time spent by individual components compared to the total when working on 20 people

from each other, it was quite surprising to see different runtimes for hands open/closed and palm/back classifiers.

4.4.2 Number of people detected

In these tests, the execution time is analyzed in relation to the number of people detected. These tests involve requesting the system to execute the full pipeline on a series of images. Each one of them has the same person repeated several times. The person used allows the

detection of all his body parts. Each image has a defined number of people, but they all have the same dimensions, therefore the unused space was filled with black pixels.



(a) (b)

Table 4.16: Examples of images used during tests

In the first test the system worked on images with a number of people between 1 and 20. The system used was operating in a docker container built starting from a Nvidia-Pytorch docker image. During tests this shown an optimization in GPU memory employed during the execution of the system, therefore it has not to be excluded the possibility that also the runtime may differ with respect to the previous test. In the first test it was registered the time elapsed between the transmission of the request from, and the receipt of, the response to the client. Results are reported in Table 4.14 and depicted in Figure 4.1. It can be seen that the runtime scales linearly with respect to the number of people detected. This constitutes a primary limitation of the system, that if not managed in the correct way could bring to unexpected results. A way to avoid this problem, could be for example, to do the preprocess to all the people detected but consider only 5 of them for the remaining pipeline components. To decide which person to consider, for instance, it can be exploited their body pose estimation confidence or their body bounding box area. However, how it can be seen in Table 4.12(c), even with only 4-5 people it is quite difficult that other people could be detected in the same image without occluding each other and therefore decreasing the total number of elements to analyze.

In the second test the same script used in the subsection 4.4.1 was employed to process the same images exploited in the previous test, registering in this time the runtime of individual components in relation to the number of people processed, however, in this case, they have been grouped by their body part of reference to have more readable results. The outcomes have been reported in Table 4.17 and depicted in Table 4.18. From Table 4.18(a) it is clear that body orientation component has the biggest impact on the runtime. From Table 4.18b instead, it can be observed better the impact of the other components. Classifiers have quite the same behavior, with those of the hands having a greater impact than those of the body, but only because four of them are used on each person rather than the two used for the body. Preprocess is the second topmost impacting runtime when only

one person is detected, but it is easily surpassed by the classifiers considered in groups when more people are present in the picture. If the PFI execution time is not considered, the head orientation is the most efficient component until around 20 people are detected.

	1	2	3	4	5	6	7	8	9	10
Preprocess	0.06505	0.06505	0.06426	0.09649	0.07269	0.07746	0.08755	0.12264	0.10257	0.11955
Head orientation	0.01318	0.02037	0.02629	0.04685	0.04458	0.05102	0.06272	0.07047	0.08309	0.08513
Body classification	0.02427	0.05176	0.07369	0.10651	0.12097	0.14289	0.17281	0.1935	0.22955	0.23738
Hand classification	0.04755	0.10359	0.14218	0.22117	0.2343	0.27552	0.33406	0.37675	0.44439	0.45783
Body orientation	0.10927	0.21924	0.31079	0.46435	0.50674	0.63688	0.69157	0.80806	0.86575	1.00425
PFI	0.00005	0.00005	0.00007	0.0001	0.00012	0.00014	0.00017	0.00019	0.00023	0.00023
	11	12	13	14	15	16	17	18	19	20
Preprocess	0.11342	0.12664	0.14684	0.15494	0.13476	0.15418	0.15055	0.16903	0.17725	0.17853
Head orientation	0.1113	0.10199	0.13113	0.1399	0.13166	0.13871	0.14725	0.18197	0.16427	0.18483
Body classification	0.26919	0.29091	0.31774	0.34931	0.36811	0.38259	0.41046	0.45087	0.45289	0.47158
Hand classification	0.55476	0.5554	0.678	0.72044	0.71199	0.74211	0.79306	0.95821	0.87865	0.9649
Body orientation	1.12086	1.20585	1.30693	1.43707	1.46156	1.53858	1.64525	1.7849	1.85469	2.00218
PFI	0.00026	0.00029	0.00031	0.00034	0.00035	0.00038	0.00041	0.00043	0.00044	0.0047

Table 4.17: Individual components total time when working on different number of people (from 1 to 20)

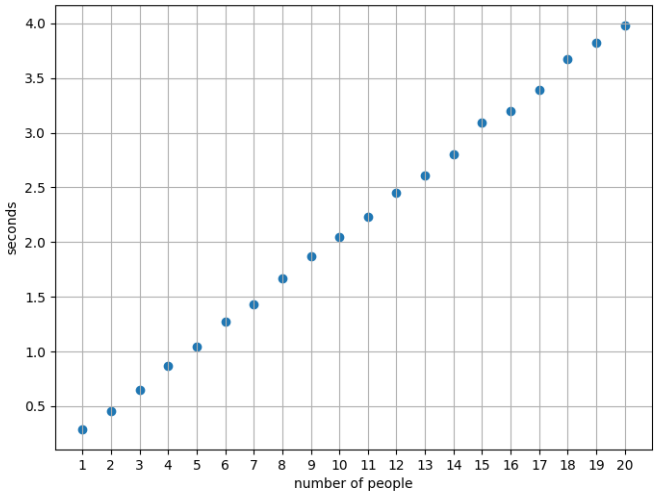
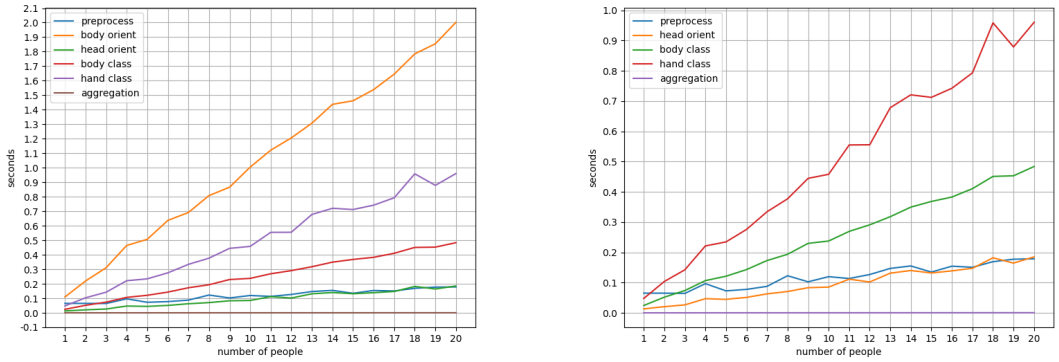


Figure 4.1: System runtime when working on different number of people (from 1 to 20)

4.4.3 Modularity and asynchronicity

Here, a test will be used to show the contribution of the modularity and asynchronicity introduced in the system. The test involves the execution of two scripts, in the first is simulate the case of a simple implementation of the system that only allows consecutive requests, therefore there would be only one client that makes the same request two times consecutively; in the second instead, the asynchronicity is tested by making two clients requesting at the same time the same individual request. The requests done to the system required to it the execution of the full pipeline over the same set of images used in tests of subsection 4.4.2. The recorded execution times started when the request was sent and stopped when the last reply have been received.



(a) (b)

Table 4.18: (a) plot of result reported in Table 4.17. (b) same as before but without plotting the body orientation values

The consecutive client employed 85.5933 seconds to finish, while the two concurrent clients took 77.6942 seconds the first and 80.3256 seconds the second. Taking the worst time between the latter two as reference, there was a decreasing in runtime of about 5 seconds while requesting the same number of processes. However, this can be effectively done on a computer with a good CPU and GPU. Indeed, a similar test conducted on a device with CPU: i5-4440 and GPU: GTX 980 with only the preprocess and body classifiers up (due to the limited amount of GPU memory), employed 23.636 seconds for the sequence, while the two concurrent clients took 25.0241 seconds the first and 25.5674 the second, employing around 1.5 seconds more requesting the same number of processes when done at the same time than when executing them in sequence.

Conclusions

A system that uses only information from a single RGB camera to estimate engagement levels in human-robot interaction has been described in this thesis.

During the definition of the system's ground idea, concepts of nonverbal communication and HRI have been reported and discussed. Having limited expertise in those fields, it was necessary to simplify the problem, redefining the meaning of engagement to be measured, as a simpler view of it: the Propensity For any type of Interaction (PFI) that a person may have toward the robot. Including the problem's requests and limitations into the equation, has brought to propose a system composed of many modules, from which, specific information would have been extracted and aggregated by an additional step to provide the final result.

During the feasibility study models able to analyze specific nonverbal cues have been tested and, in the end, pose estimation, head orientation, body orientation and the classification of body and hands poses, have been chosen as the information to be extracted by the system modules.

For the first three components, already built API or models were found, for the remaining instead, custom datasets have been annotated and fine-tuning has been used to train specific classifiers.

The system has been tested in some specific scenarios as well as more general ones, showing good accuracies in almost all modules and great final results but highlighting at the same time a serious problem for what regards the hands' modules, that as maintained, it could be caused by the dataset composed for those tasks. Other tests documented some limitations of the system, however, the specific cases in which they may appear have been described along with some improvements to make in order to avoid those strange behaviors. Finally, some runtime tests have highlighted that when many people are completely visible, the execution time of the system could be unfeasible with the possible usage described during chapters. Some solutions have been proposed to overcome the problem, however, it has also been shown that such situations may be difficult to appear. At the same time, those tests have shown that the system modular design combined with the asynchronicity introduced have reduced the time spent for requests by almost 11% with respect to a simple consecutive pipeline.

Surely, other tests should be conducted to confirm the goodness of the system and to verify its consistency, however, as shown so far, the system created can be considered a good start moreover if evaluated as an explorative or initial solution for the problem analyzed. Nevertheless, it still has a lot of room for improvement, for example, better model training procedures or the use of more recent methods can enhance each defined module. Also, other modules can be added, not only to extract other relevant information but also to process the already present ones, and have in this way, more to aggregate in the final step. For instance, by going to keeping track of the already present information, changes in particular aspects can be detected, and therefore, specific nonverbal cues can be exploited.

Another thing to keep in mind is that the datasets used and the models that were already built were all published under non-commercial uses licenses, therefore, if someone wants to exploit the system created in any commercial products, it has to re-create them from scratch or buy their commercial licenses if there are any.

Appendix

Hyperparameters study tables

Trial	Batch Size	Learning Rate	Optimizer	Rcs low	Accurcay
1	64	0.000252	SGD	0.5	0.936
2	16	0.0107	SGD	0.6	0.962
3	32	0.0000354	Adam	0.6	0.96
4	32	0.035	SGD	0.8	0.951
5	16	0.0000424	SGD	0.7	0.953
6	64	0.0000786	Adam	0.5	0.962
7	64	0.0544	SGD	0.5	0.951
8	16	0.00016	SGD	0.7	0.96
9	32	0.000463	SGD	0.6	0.96
10	16	0.0000189	Adam	0.6	0.961
11	64	0.00294	Adam	0.9	0.907
12	16	0.00355	Adam	0.5	0.894
13	64	0.00843	Adam	0.8	0.581
14	64	0.0123	Adam	0.9	0.581
15	16	0.00113	Adam	0.6	0.903
16	16	0.0000794	SGD	0.5	0.936
17	64	0.00107	SGD	0.6	0.96
18	64	0.0000104	Adam	0.5	0.962
19	64	0.0000109	Adam	0.5	0.901
20	16	0.0206	SGD	0.8	0.955
21	32	0.00305	SGD	0.7	0.963
22	32	0.00342	SGD	0.7	0.961
23	32	0.094	SGD	0.7	0.818
24	32	0.00747	SGD	0.7	0.959
25	32	0.00175	SGD	0.9	0.955
26	16	0.000446	SGD	0.7	0.959
27	32	0.00605	Adam	0.6	0.907
28	64	0.0209	SGD	0.5	0.962
29	64	0.0237	Adam	0.5	0.581
30	64	0.0000806	SGD	0.5	0.885
31	64	0.000308	SGD	0.5	0.962
32	64	0.000302	SGD	0.5	0.961
33	64	0.0002	SGD	0.5	0.958
34	64	0.000621	SGD	0.5	0.965
35	64	0.000719	SGD	0.5	0.96
36	64	0.00166	SGD	0.5	0.961
37	32	0.000571	SGD	0.8	0.956
38	64	0.000133	SGD	0.5	0.958
39	64	0.0514	SGD	0.7	0.949
40	32	0.0000306	Adam	0.5	0.966
41	32	0.0000362	Adam	0.7	0.966
44	32	0.0000442	Adam	0.7	0.958
45	32	0.0000233	Adam	0.7	0.955
46	32	0.0000642	Adam	0.7	0.962
47	32	0.0000245	Adam	0.9	0.961
48	32	0.000278	Adam	0.7	0.95
49	32	0.0000408	Adam	0.8	0.939
50	32	0.000133	Adam	0.5	0.953
51	32	0.0136	SGD	0.7	0.957
52	32	0.000017	SGD	0.5	0.942
53	32	0.00158	Adam	0.9	0.581
54	64	0.000102	Adam	0.5	0.954
55	64	0.0000494	Adam	0.5	0.965
56	64	0.0000539	Adam	0.5	0.967
57	64	0.0000507	Adam	0.5	0.962
58	64	0.0000286	Adam	0.5	0.961
59	32	0.0000571	Adam	0.6	0.945
60	64	0.0000158	Adam	0.7	0.961
61	16	0.0000374	Adam	0.5	0.963
62	16	0.000013	Adam	0.5	0.964
63	16	0.0000142	Adam	0.5	0.962
65	16	0.0000267	Adam	0.5	0.961
66	16	0.000039	Adam	0.5	0.96
67	16	0.0000101	Adam	0.5	0.962
68	32	0.00268	Adam	0.8	0.581
69	16	0.00456	Adam	0.7	0.887
70	64	0.0000845	Adam	0.5	0.964
71	64	0.0000943	Adam	0.5	0.96
72	64	0.0000205	Adam	0.5	0.963
73	64	0.0000302	Adam	0.5	0.964
74	64	0.0000309	Adam	0.5	0.961
75	64	0.0000623	Adam	0.5	0.968
76	64	0.000065	Adam	0.5	0.96
77	64	0.0000141	Adam	0.5	0.961
78	64	0.00002	Adam	0.5	0.96
79	64	0.000196	Adam	0.6	0.951
80	64	0.0000797	Adam	0.5	0.963
81	64	0.00013	Adam	0.9	0.958
82	64	0.0000514	Adam	0.5	0.888
83	64	0.0000347	Adam	0.5	0.963
84	64	0.000109	Adam	0.5	0.959
85	64	0.0000623	Adam	0.5	0.96
88	64	0.0000301	Adam	0.5	0.961
89	64	0.0000121	Adam	0.5	0.958
90	64	0.0000219	Adam	0.5	0.957
92	16	0.000176	Adam	0.5	0.944
93	64	0.0000475	Adam	0.8	0.958
94	64	0.0000727	Adam	0.5	0.96
95	64	0.000355	Adam	0.5	0.94
96	16	0.0000171	Adam	0.6	0.96
97	32	0.00087	Adam	0.5	0.927
98	64	0.0000757	Adam	0.5	0.959
99	64	0.0000437	Adam	0.5	0.962
100	64	0.0000848	Adam	0.5	0.96
101	64	0.0000339	Adam	0.5	0.959
102	64	0.0000568	Adam	0.9	0.961
105	64	0.000108	Adam	0.5	0.956
106	32	0.00022	Adam	0.5	0.952
107	64	0.00015	Adam	0.5	0.956
108	64	0.0000244	SGD	0.5	0.935
110	16	0.0000127	Adam	0.8	0.96
112	64	0.0000186	Adam	0.5	0.966
113	64	0.000027	Adam	0.5	0.964
114	64	0.0000179	Adam	0.5	0.962
115	64	0.0000278	Adam	0.5	0.964
116	32	0.0000392	Adam	0.7	0.962
117	64	0.000027	Adam	0.5	0.962
118	64	0.0000223	Adam	0.5	0.963
119	64	0.0000326	SGD	0.5	0.942
120	32	0.0000141	Adam	0.6	0.925
121	64	0.0000483	Adam	0.5	0.963
122	64	0.0000196	Adam	0.5	0.963
123	64	0.0000281	Adam	0.5	0.962
124	64	0.0000342	Adam	0.5	0.963
125	64	0.0000162	Adam	0.5	0.967
126	64	0.0000156	Adam	0.7	0.961
127	16	0.0000115	Adam	0.5	0.963
128	64	0.0000228	Adam	0.5	0.963
129	32	0.0000101	SGD	0.9	0.915
130	64	0.0000172	Adam	0.5	0.964
131	64	0.0000404	Adam	0.5	0.963
132	64	0.0000171	Adam	0.5	0.964
133	64	0.0000586	Adam	0.5	0.961
134	64	0.0000277	Adam	0.5	0.962
135	64	0.0000187	Adam	0.5	0.963
136	64	0.0000164	Adam	0.5	0.965
137	64	0.0000446	Adam	0.5	0.964
138	16	0.0000126	Adam	0.7	0.958
139	64	0.0000496	Adam	0.5	0.963
140	64	0.0000182	SGD	0.5	0.79
141	64	0.0000392	Adam	0.8	0.96
142	64	0.0000157	Adam	0.5	0.963
143	64	0.0000134	Adam	0.5	0.963
144	32	0.0000222	Adam	0.5	0.961
145	64	0.000029	Adam	0.5	0.961
146	64	0.0000204	Adam	0.5	0.96
147	64	0.0000151	Adam	0.5	0.962
148	32	0.0000245	Adam	0.5	0.962
149	64	0.0000355	Adam	0.6	0.96
150	16	0.0000643	Adam	0.5	0.961

Table 4.19: Results of the hyperparameter study for the upstanding/sitting/lying classifier

Trial	Batch Size	Learning Rate	Optimizer	Rcs Low	Accuracy	Trial	Batch Size	Learning Rate	Optimizer	Rcs Low	Accuracy
1	64	0.0000142	SGD	0.5	0.642	50	32	0.0000112	SGD	0.08	0.622
2	32	0.00101	SGD	0.5	0.853	51	64	0.00325	SGD	0.5	0.863
3	64	0.00131	Adam	0.9	0.679	52	64	0.00488	SGD	0.5	0.864
4	64	0.00411	SGD	0.5	0.86	53	64	0.00283	SGD	0.5	0.857
5	32	0.0304	Adam	0.5	0.551	54	64	0.00153	SGD	0.5	0.852
6	32	0.00912	Adam	0.08	0.551	55	64	0.0209	SGD	0.5	0.858
7	32	0.0369	SGD	0.08	0.861	56	64	0.00797	SGD	0.08	0.852
8	64	0.0991	Adam	0.9	0.551	57	32	0.00415	SGD	0.5	0.861
9	64	0.0571	Adam	0.5	0.551	58	64	0.00121	SGD	0.5	0.842
10	64	0.0000146	SGD	0.5	0.697	59	32	0.00392	SGD	0.5	0.86
11	32	0.000122	SGD	0.08	0.796	60	32	0.00215	SGD	0.5	0.862
12	32	0.00647	SGD	0.08	0.853	61	32	0.00233	SGD	0.5	0.86
13	64	0.0069	SGD	0.08	0.852	62	32	0.00288	SGD	0.5	0.862
14	32	0.0182	SGD	0.08	0.851	63	32	0.0032	SGD	0.5	0.855
15	64	0.00303	SGD	0.9	0.851	64	32	0.00197	SGD	0.5	0.866
16	32	0.000404	SGD	0.5	0.836	65	32	0.0017	SGD	0.5	0.854
17	64	0.000227	SGD	0.08	0.786	66	32	0.000914	SGD	0.5	0.846
18	32	0.022	SGD	0.08	0.86	67	32	0.00232	SGD	0.5	0.857
19	32	0.0227	SGD	0.08	0.854	68	64	0.000537	SGD	0.5	0.834
20	64	0.00217	SGD	0.5	0.857	69	32	0.00113	Adam	0.5	0.598
21	32	0.0472	SGD	0.08	0.844	70	32	0.00521	SGD	0.5	0.861
22	32	0.0115	SGD	0.08	0.863	71	64	0.000359	SGD	0.5	0.832
23	32	0.0105	SGD	0.08	0.854	72	32	0.0057	SGD	0.5	0.856
24	32	0.00228	SGD	0.9	0.852	73	32	0.00181	SGD	0.5	0.853
25	32	0.00431	SGD	0.08	0.852	74	32	0.00428	SGD	0.5	0.857
26	32	0.0121	SGD	0.5	0.86	75	32	0.0027	SGD	0.5	0.85
27	32	0.0762	Adam	0.08	0.551	76	32	0.00983	SGD	0.5	0.858
28	64	0.000657	SGD	0.08	0.823	77	32	0.00728	SGD	0.08	0.863
29	64	0.0000637	SGD	0.5	0.77	80	32	0.00764	SGD	0.08	0.857
30	32	0.0163	SGD	0.9	0.86	81	32	0.0032	SGD	0.08	0.852
31	32	0.0363	SGD	0.9	0.847	82	64	0.0169	Adam	0.5	0.551
32	32	0.0203	SGD	0.9	0.85	83	32	0.0000226	SGD	0.5	0.761
33	32	0.00508	SGD	0.9	0.856	84	32	0.00473	SGD	0.5	0.861
34	32	0.0136	SGD	0.9	0.847	85	32	0.00536	SGD	0.5	0.859
35	32	0.00155	SGD	0.5	0.859	86	32	0.00328	SGD	0.08	0.849
36	32	0.0343	SGD	0.08	0.848	87	32	0.0256	SGD	0.08	0.854
37	64	0.00377	Adam	0.08	0.551	88	32	0.0022	SGD	0.08	0.854
38	32	0.0134	SGD	0.9	0.85	89	64	0.0109	SGD	0.08	0.86
39	64	0.00839	Adam	0.5	0.551	90	32	0.00879	SGD	0.5	0.861
40	32	0.0652	SGD	0.08	0.661	91	32	0.00776	SGD	0.5	0.86
41	32	0.0271	Adam	0.08	0.551	92	32	0.00185	SGD	0.5	0.856
42	32	0.045	SGD	0.08	0.65	93	64	0.000991	SGD	0.5	0.844
43	32	0.0985	SGD	0.5	0.574	94	32	0.00678	SGD	0.08	0.858
44	32	0.0136	SGD	0.5	0.853	95	32	0.0155	SGD	0.08	0.857
45	64	0.00597	SGD	0.5	0.86	96	32	0.00436	SGD	0.5	0.859
46	64	0.00638	SGD	0.5	0.86	97	32	0.0429	SGD	0.08	0.828
47	64	0.000758	SGD	0.5	0.838	98	32	0.00266	SGD	0.5	0.854
48	32	0.00916	SGD	0.5	0.863	99	32	0.00507	Adam	0.5	0.551
49	32	0.0105	Adam	0.9	0.551	100	32	0.00142	SGD	0.5	0.85

Table 4.20: Results of the hyperparameter study for the body open/closed classifier

Trial	Batch Size	Learning Rate	Optimizer	Rcs low	Accuracy	Trial	Batch Size	Learning Rate	Optimizer	Rcs low	Accuracy
1	32	0.000023	Adam	0.5	0.867	52	64	0.00142	Adam	0.5	0.519
2	32	0.000684	SGD	0.5	0.847	53	64	0.0000194	Adam	0.5	0.862
3	64	0.00381	Adam	0.08	0.519	54	32	0.0000356	Adam	0.5	0.87
5	64	0.0681	SGD	0.08	0.519	55	32	0.0000384	Adam	0.5	0.859
6	32	0.00924	SGD	0.9	0.865	56	32	0.0000139	Adam	0.5	0.861
7	64	0.000918	SGD	0.9	0.842	57	32	0.0925	Adam	0.5	0.519
8	32	0.0000128	Adam	0.5	0.868	58	32	0.0000101	Adam	0.5	0.863
9	32	0.0000539	SGD	0.08	0.733	59	32	0.000022	Adam	0.5	0.865
10	32	0.000052	SGD	0.5	0.751	60	32	0.0000745	Adam	0.08	0.855
11	64	0.000258	SGD	0.9	0.794	61	32	0.0000273	SGD	0.5	0.692
12	32	0.0000107	Adam	0.5	0.864	62	32	0.0000504	Adam	0.9	0.863
13	32	0.000013	Adam	0.5	0.86	63	32	0.0000144	Adam	0.5	0.87
14	32	0.0000697	Adam	0.5	0.857	64	32	0.00464	Adam	0.5	0.519
15	32	0.000163	Adam	0.5	0.847	65	32	0.0000143	Adam	0.5	0.867
16	32	0.0000228	Adam	0.5	0.866	66	32	0.0000311	Adam	0.5	0.868
17	32	0.000215	Adam	0.5	0.841	67	32	0.0000211	Adam	0.5	0.857
18	32	0.000032	Adam	0.5	0.862	68	32	0.0000151	Adam	0.5	0.865
20	32	0.00406	Adam	0.5	0.519	69	32	0.0000344	Adam	0.5	0.867
21	64	0.000127	Adam	0.9	0.858	70	32	0.00012	Adam	0.5	0.853
22	32	0.000011	Adam	0.08	0.857	71	32	0.0000507	Adam	0.5	0.86
23	32	0.000465	Adam	0.5	0.811	72	64	0.000714	Adam	0.5	0.641
24	32	0.0000251	Adam	0.5	0.869	73	32	0.0000305	Adam	0.08	0.857
25	32	0.000024	Adam	0.5	0.859	74	32	0.0000191	Adam	0.5	0.859
26	32	0.0000928	Adam	0.5	0.854	75	32	0.0000136	Adam	0.5	0.862
27	32	0.0000246	Adam	0.5	0.861	76	32	0.000024	Adam	0.5	0.866
28	32	0.0000377	Adam	0.5	0.863	77	32	0.0000101	Adam	0.5	0.866
29	64	0.00197	Adam	0.5	0.519	78	32	0.0000395	Adam	0.5	0.863
30	32	0.000371	Adam	0.5	0.828	79	32	0.000212	SGD	0.9	0.824
31	32	0.00009	Adam	0.9	0.857	80	64	0.0000218	Adam	0.5	0.872
32	32	0.0331	Adam	0.08	0.519	81	64	0.0000692	Adam	0.5	0.858
33	32	0.0000185	Adam	0.5	0.866	82	64	0.0000133	Adam	0.5	0.865
34	32	0.0000175	Adam	0.5	0.867	83	64	0.0000463	Adam	0.5	0.862
35	32	0.0000418	Adam	0.5	0.86	84	64	0.0000244	Adam	0.5	0.867
36	32	0.0000161	Adam	0.5	0.869	85	64	0.0000169	Adam	0.5	0.863
38	32	0.0000162	Adam	0.5	0.862	86	64	0.0000293	Adam	0.5	0.867
39	64	0.0000158	Adam	0.5	0.869	87	32	0.00002	Adam	0.5	0.87
40	64	0.0000111	SGD	0.08	0.583	88	64	0.0000187	Adam	0.5	0.86
41	64	0.0000397	Adam	0.5	0.866	89	32	0.0000342	Adam	0.5	0.862
42	64	0.0000874	SGD	0.9	0.733	90	64	0.0000123	Adam	0.08	0.858
43	64	0.0000285	Adam	0.08	0.865	91	32	0.0000206	SGD	0.5	0.647
44	64	0.0000555	SGD	0.5	0.71	92	32	0.0000622	Adam	0.5	0.864
45	64	0.0000198	Adam	0.5	0.871	93	64	0.000017	Adam	0.9	0.877
46	64	0.000016	Adam	0.5	0.868	94	64	0.0000164	Adam	0.9	0.871
47	64	0.0174	Adam	0.5	0.519	95	64	0.0000161	Adam	0.9	0.869
48	64	0.0000585	Adam	0.5	0.865	96	64	0.0000152	Adam	0.9	0.863
49	64	0.0000113	Adam	0.5	0.866	97	64	0.0000244	Adam	0.9	0.868
50	64	0.0000284	SGD	0.9	0.619	98	64	0.0000181	Adam	0.9	0.876
51	64	0.000136	Adam	0.5	0.85	99	64	0.0000118	Adam	0.9	0.87
						100	64	0.0000121	Adam	0.9	0.863

Table 4.21: Results of the hyperparameter study for the hands palm/back classifier

Trial	Batch Size	Learning Rate	Optimizer	Rcs low	Accuracy	Trial	Batch Size	Learning Rate	Optimizer	Rcs low	Accuracy
1	64	0.015	Adam	0.5	0.547	50	64	0.0000633	Adam	0.08	0.883
2	32	0.0000872	SGD	0.9	0.836	51	32	0.000125	Adam	0.9	0.887
3	64	0.00424	Adam	0.08	0.547	52	64	0.0000115	Adam	0.5	0.889
5	64	0.0000616	SGD	0.08	0.777	53	64	0.000036	Adam	0.5	0.899
6	64	0.011	SGD	0.9	0.885	54	64	0.0000289	Adam	0.5	0.892
7	64	0.000962	SGD	0.5	0.873	55	64	0.0000234	Adam	0.5	0.897
8	64	0.00375	SGD	0.5	0.891	56	64	0.0000256	Adam	0.5	0.897
9	32	0.0163	SGD	0.9	0.882	57	64	0.000029	Adam	0.5	0.894
10	64	0.000244	Adam	0.08	0.883	58	64	0.000028	Adam	0.5	0.896
11	32	0.000532	SGD	0.08	0.857	59	64	0.0000274	Adam	0.5	0.9
12	32	0.00238	Adam	0.5	0.547	60	64	0.000037	Adam	0.5	0.894
13	64	0.0962	SGD	0.9	0.547	61	32	0.0000176	Adam	0.5	0.894
14	64	0.0152	SGD	0.5	0.891	62	64	0.000167	Adam	0.5	0.897
15	64	0.0708	SGD	0.5	0.696	63	64	0.000135	Adam	0.5	0.891
16	64	0.0000112	SGD	0.5	0.746	66	64	0.0000452	Adam	0.5	0.893
17	64	0.00349	SGD	0.5	0.888	67	64	0.000018	Adam	0.5	0.898
18	64	0.0255	SGD	0.5	0.889	68	64	0.0000198	Adam	0.5	0.897
19	64	0.00591	SGD	0.5	0.887	69	64	0.0000158	Adam	0.5	0.895
20	32	0.0312	Adam	0.5	0.547	70	64	0.000122	Adam	0.5	0.887
21	64	0.0019	SGD	0.5	0.886	71	64	0.0000505	Adam	0.5	0.892
22	64	0.00105	SGD	0.5	0.886	72	64	0.0000226	Adam	0.08	0.893
23	64	0.0454	SGD	0.5	0.881	73	64	0.000202	Adam	0.9	0.876
24	64	0.00887	SGD	0.5	0.894	74	64	0.0000137	Adam	0.5	0.895
25	64	0.00777	SGD	0.5	0.89	75	64	0.0000194	Adam	0.5	0.897
26	64	0.00968	SGD	0.5	0.891	76	64	0.0000103	Adam	0.5	0.899
27	64	0.0223	SGD	0.5	0.891	77	64	0.0000108	Adam	0.5	0.899
28	32	0.00895	Adam	0.5	0.547	78	64	0.0000102	Adam	0.5	0.892
29	64	0.046	SGD	0.5	0.877	79	64	0.0000222	Adam	0.5	0.894
30	64	0.00128	SGD	0.08	0.865	81	64	0.0000142	Adam	0.5	0.895
31	64	0.0136	Adam	0.9	0.547	82	32	0.0000379	Adam	0.5	0.897
32	64	0.012	SGD	0.5	0.889	83	32	0.0000376	Adam	0.5	0.897
33	64	0.0237	SGD	0.5	0.886	84	32	0.0000451	Adam	0.5	0.894
34	64	0.0477	SGD	0.5	0.865	85	32	0.0000346	Adam	0.5	0.896
35	64	0.017	SGD	0.5	0.889	86	32	0.0000129	Adam	0.5	0.902
36	64	0.00691	SGD	0.5	0.89	87	32	0.0000132	Adam	0.5	0.899
37	64	0.00568	SGD	0.08	0.874	88	32	0.0000126	Adam	0.5	0.897
38	32	0.00243	SGD	0.9	0.88	89	32	0.0000169	Adam	0.5	0.897
39	64	0.000449	Adam	0.5	0.876	90	32	0.0000106	Adam	0.5	0.898
40	64	0.0237	SGD	0.5	0.891	91	32	0.000011	Adam	0.08	0.892
41	64	0.00425	SGD	0.9	0.887	92	32	0.0000186	Adam	0.9	0.891
42	32	0.0335	SGD	0.08	0.882	93	32	0.0000142	Adam	0.5	0.895
43	64	0.0209	SGD	0.5	0.889	94	32	0.0000102	Adam	0.5	0.897
44	64	0.0105	SGD	0.5	0.888	95	32	0.000353	Adam	0.5	0.879
45	64	0.00346	SGD	0.5	0.888	96	32	0.0000258	Adam	0.5	0.897
46	64	0.0147	SGD	0.5	0.891	97	32	0.0000161	Adam	0.5	0.897
47	64	0.0667	SGD	0.5	0.547	98	32	0.0000165	Adam	0.5	0.895
48	64	0.0000807	Adam	0.5	0.897	99	32	0.0000239	Adam	0.5	0.892
49	64	0.0000771	Adam	0.5	0.895	100	64	0.0000125	Adam	0.5	0.894

Table 4.22: Results of the hyperparameter study for the hands open/closed classifier

Bibliography

[1] Natasha Abner, Kensy Cooperrider, and Susan Goldin-Meadow. “Gesture for Linguists: A Handy Primer”. In: *Language and Linguistics Compass* 9.11 (2015), pp. 437–451. DOI: https://doi.org/10.1111/lnc3.12168. eprint: https://compass.onlinelibrary.wiley.com/doi/pdf/10.1111/lnc3.12168. URL: https://compass.onlinelibrary.wiley.com/doi/abs/10.1111/lnc3.12168.

[2] Reginald B Adams Jr and Robert E Kleck. “Effects of direct and averted gaze on the perception of facially communicated emotion.” In: *Emotion* 5.1 (2005), p. 3.

[3] Reginald B Adams Jr and Robert E Kleck. “Perceived gaze direction and the processing of facial displays of emotion”. In: *Psychological science* 14.6 (2003), pp. 644–647.

[4] Ralph Adolphs et al. “A mechanism for impaired fear recognition after amygdala damage”. In: *Nature* 433.7021 (2005), pp. 68–72.

[5] John R Aiello. “Human spatial behavior”. In: *Handbook of environmental psychology* 1.1987 (1987), pp. 389–504.

[6] Takuya Akiba et al. “Optuna: A Next-generation Hyperparameter Optimization Framework”. In: *CoRR* abs/1907.10902 (2019). arXiv: 1907.10902. URL: http://arxiv.org/abs/1907.10902.

[7] Tony Alessandra. *Charisma: seven keys to developing the magnetism that leads to success*. Business Plus, 2000.

[8] Philipp Althaus et al. “Navigation for human-robot interaction tasks”. In: *IEEE International Conference on Robotics and Automation, 2004. Proceedings. ICRA’04. 2004*. Vol. 2. IEEE. 2004, pp. 1894–1900.

[9] Irwin Altman. “The environment and social behavior: privacy, personal space, territory, and crowding.” In: (1975).

[10] Janis F Andersen, Peter A Andersen, and Myron W Lustig. “Opposite sex touch avoidance: A national replication and extension”. In: *Journal of nonverbal behavior* 11.2 (1987), pp. 89–109.

[11] Peter A Andersen. “Nonverbal immediacy in interpersonal communication”. In: *Multichannel integrations of nonverbal behavior* (1985), pp. 1–36.

[12] Peter A Andersen. “Tactile traditions: Cultural differences and similarities in haptic communication”. In: *The handbook of touch: Neuroscience, behavioral, and health perspectives* (2011), pp. 351–369.

[13] Peter A Andersen and Laura K Guerrero. “Haptic behavior in social interaction”. In: *Human haptic perception: Basics and applications*. Springer, 2008, pp. 155–163.

[14] Peter A Andersen and Kenneth Leibowitz. “The development and nature of the construct touch avoidance”. In: *Environmental psychology and nonverbal behavior* 3.2 (1978), pp. 89–106.

[15] Mykhaylo Andriluka, Stefan Roth, and Bernt Schiele. “Monocular 3d pose estimation and tracking by detection”. In: *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. Ieee. 2010, pp. 623–630.

[16] Sean Andrist et al. “Conversational gaze aversion for humanlike robots”. In: *2014 9th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE. 2014, pp. 25–32.

[17] Morris Antonello et al. “Fast and Robust detection of fallen people from a mobile robot”. In: *Intelligent Robots and Systems (IROS), 2017 IEEE/RSJ International Conference on*. IEEE. 2017.

[18] Salvatore M Anzalone et al. “Evaluating the engagement with social robots”. In: *International Journal of Social Robotics* 7.4 (2015), pp. 465–478.

[19] Robert Ardrey. “The territorial imperative: A personal inquiry into the animal origins of property and nations.” In: (1966).

[20] Brenna D Argall et al. “A survey of robot learning from demonstration”. In: *Robotics and autonomous systems* 57.5 (2009), pp. 469–483.

[21] Michael Argyle and Mark Cook. “Gaze and mutual gaze.” In: (1976).

[22] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. “Layer normalization”. In: *arXiv preprint arXiv:1607.06450* (2016).

[23] Davide Baltieri, Roberto Vezzani, and Rita Cucchiara. “3dpes: 3d people dataset for surveillance and forensics”. In: *Proceedings of the 2011 joint ACM workshop on Human gesture and behavior understanding*. 2011, pp. 59–64.

[24] Davide Baltieri, Roberto Vezzani, and Rita Cucchiara. “People orientation recognition by mixtures of wrapped distributions on random trees”. In: *European conference on computer vision*. Springer. 2012, pp. 270–283.

[25] Simon Baron-Cohen et al. “Is there a” language of the eyes”? Evidence from normal adults, and adults with autism or Asperger syndrome”. In: *Visual cognition* 4.3 (1997), pp. 311–331.

[26] Godfrey T Barrett-Lennard. “Dimensions of therapist response as causal factors in therapeutic change.” In: *Psychological monographs: General and applied* 76.43 (1962), p. 1.

[27] Janet Beavin Bavelas and Nicole Chovil. “Nonverbal and Verbal Communication: Hand Gestures and Facial Displays as Part of Language Use in Face-to-face Dialogue.” In: (2006).

[28] Paul Baxter et al. “Tracking gaze over time in HRI as a proxy for engagement and attribution of social agency”. In: *Proceedings of the 2014 ACM/IEEE international conference on Human-robot interaction*. 2014, pp. 126–127.

[29] Jackson Beatty, Brennis Lucero-Wagoner, et al. “The pupillary system”. In: *Handbook of psychophysiology* 2.142-162 (2000).

[30] Aryel Beck et al. “Interpretation of emotional body language displayed by robots”. In: *Proceedings of the 3rd international workshop on Affective interaction in natural environments*. 2010, pp. 37–42.

[31] Irwan Bello et al. “Revisiting resnets: Improved training and scaling strategies”. In: *Advances in Neural Information Processing Systems* 34 (2021), pp. 22614–22627.

[32] Tony Belpaeme et al. “Multimodal child-robot interaction: Building social bonds”. In: *Journal of Human-Robot Interaction* 1.2 (2012).

[33] Atef Ben-Youssef et al. “UE-HRI: a new dataset for the study of user engagement in spontaneous human-robot interactions”. In: *Proceedings of the 19th ACM international conference on multimodal interaction*. 2017, pp. 464–472.

[34] Rodrigo Benenson, Stefan Popov, and Vittorio Ferrari. “Large-scale interactive object segmentation with human annotators”. In: *CVPR*. 2019.

[35] Christopher P Benton. “Rapid reactions to direct and averted facial expressions of fear and anger”. In: *Visual Cognition* 18.9 (2010), pp. 1298–1319.

[36] James Bergstra, Daniel Yamins, and David Cox. “Making a science of model search: Hyperparameter optimization in hundreds of dimensions for vision architectures”. In: *International conference on machine learning*. PMLR. 2013, pp. 115–123.

[37] James Bergstra et al. “Algorithms for hyper-parameter optimization”. In: *Advances in neural information processing systems* 24 (2011).

[38] Michael H Bond and Daisuke Shiraishi. “The effect of body lean and status of an interviewer on the non-verbal behavior of Japanese interviewees”. In: *International Journal of Psychology* 9.2 (1974), pp. 117–128.

[39] G. Bradski. “The OpenCV Library”. In: *Dr. Dobb’s Journal of Software Tools* (2000).

[40] Cynthia Breazeal. *Designing sociable robots*. MIT press, 2004.

[41] Paul Bremner et al. “Conversational gestures in human-robot interaction”. In: *2009 IEEE international conference on systems, man and cybernetics*. IEEE. 2009, pp. 1645–1649.

[42] Dražen Brščić et al. “Person tracking in large public spaces using 3-D range sensors”. In: *IEEE Transactions on Human-Machine Systems* 43.6 (2013), pp. 522–534.

[43] Adrian Bulat and Georgios Tzimiropoulos. “How far are we from solving the 2d & 3d face alignment problem?(and a dataset of 230,000 3d facial landmarks)”. In: *Proceedings of the IEEE International Conference on Computer Vision*. 2017, pp. 1021–1030.

[44] P. E. Bull. *The communication of emotion*. Paper presented at the annual conference of the British Psychological Society, University of Warwick. 1984.

[45] Peter Bull. *Body movement and interpersonal communication*. Chichester ; New York : Wiley, 1983.

[46] Peter E Bull. “Individual differences in non-verbal communication”. In: *Individual differences in movement*. Springer, 1985, pp. 231–245.

[47] Peter E Bull. *Posture & gesture*. Vol. 16. Elsevier, 2016.

[48] Judee K Burgoon. “Relational message interpretations of touch, conversational distance, and posture”. In: *Journal of Nonverbal behavior* 15.4 (1991), pp. 233–259.

[49] Judee K Burgoon, Kory Floyd, and Laura K Guerrero. “Nonverbal communication theories of interpersonal adaptation”. In: *The new SAGE handbook of communication science*. Sage, 2010, pp. 93–110.

[50] Judee K Burgoon and Thomas Saine. *The unspoken dialogue: An introduction to nonverbal communication*. Houghton Mifflin School, 1978.

[51] Xavier P Burgos-Artizzu et al. “Merging pose estimates across space and time”. In: (2013).

[52] Maya Cakmak et al. “Human preferences for robot-human hand-over configurations”. In: *2011 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE. 2011, pp. 1986–1993.

[53] John B Calhoun. “Population density and social pathology”. In: *Scientific American* 206.2 (1962), pp. 139–149.

[54] John Canny. “A Computational Approach to Edge Detection”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* PAMI-8.6 (1986), pp. 679–698. DOI: 10.1109/TPAMI.1986.4767851.

[55] Dana R Carney, Judith A Hall, and Lavonia Smith LeBeau. “Beliefs about the nonverbal expression of social power”. In: *Journal of Nonverbal Behavior* 29.2 (2005), pp. 105–123.

[56] Barbara A Cavallin and B Kent Houston. “Aggressiveness, maladjustment, body experience and the protective function of personal space”. In: *Journal of Clinical Psychology* 36.1 (1980), pp. 170–176.

[57] Michael RA Chance. “An interpretation of some agonistic postures; the role of “cut-off” acts and postures”. In: *Symposium of the Zoological Society of London*. Vol. 8. 71. 1962, p. 57.

[58] Feng-Ju Chang et al. “Faceposenet: Making a case for landmark-free face alignment”. In: *Proceedings of the IEEE International Conference on Computer Vision Workshops*. 2017, pp. 1599–1608.

[59] Cheng Chen, Alexandre Heili, and Jean-Marc Odobez. “Combined estimation of location and body pose in surveillance video”. In: *2011 8th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*. IEEE. 2011, pp. 5–10.

[60] Xianjie Chen and Alan L Yuille. “Parsing occluded people by flexible compositions”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2015, pp. 3945–3954.

[61] Jinyoung Choi, Beom-Jin Lee, and Byoung-Tak Zhang. “Human body orientation estimation using convolutional neural network”. In: *arXiv preprint arXiv:1609.01984* (2016).

[62] John J Christian. *Phenomena associated with population density*. 1961.

[63] Dennis Coon and John O Mitterer. *Introduction to Psychology: Gateways to Mind and Behavior*. Wadsworth Cengage Learning, 2010.

[64] Martin Cooney et al. “Designing enjoyable motion-based play interactions with a small humanoid robot”. In: *International Journal of Social Robotics* 6.2 (2014), pp. 173–193.

[65] Timothy F. Cootes, Gareth J. Edwards, and Christopher J. Taylor. “Active appearance models”. In: *IEEE Transactions on pattern analysis and machine intelligence* 23.6 (2001), pp. 681–685.

[66] Mark Coulson. “Attributing Emotion to Static Body Postures: Recognition Accuracy, Confusions, and Viewpoint Dependence”. In: *Journal of Nonverbal Behavior* 28.2 (June 2004), pp. 117–139. ISSN: 1573-3653. DOI: 10.1023/B:JONB.0000023655.25550.be. URL: https://doi.org/10.1023/B:JONB.0000023655.25550.be.

[67] Ekin D Cubuk et al. “Randaugment: Practical automated data augmentation with a reduced search space”. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*. 2020, pp. 702–703.

[68] M. Danesi. “Kinesics”. In: *Encyclopedia of Language & Linguistics (Second Edition)*. Ed. by Keith Brown. Second Edition. Oxford: Elsevier, 2006, pp. 207–213. ISBN: 978-0-08-044854-1. DOI: https://doi.org/10.1016/B0-08-044854-2/01421-8. URL: https://www.sciencedirect.com/science/article/pii/B0080448542014218.

[69] Charles Darwin and Phillip Prodger. *The expression of the emotions in man and animals*. Oxford University Press, USA, 1998.

[70] Andrew J Davison et al. “MonoSLAM: Real-time single camera SLAM”. In: *IEEE transactions on pattern analysis and machine intelligence* 29.6 (2007), pp. 1052–1067.

[71] Jia Deng et al. “Imagenet: A large-scale hierarchical image database”. In: *2009 IEEE conference on computer vision and pattern recognition*. Ieee. 2009, pp. 248–255.

[72] Bella M DePaulo and Howard S Friedman. “Nonverbal communication.” In: (1998).

[73] Felix Deutsch. “Analysis of postural behavior”. In: *The Psychoanalytic Quarterly* 16.2 (1947), pp. 195–213.

[74] Felix Deutsch. “Analytic posturology”. In: *The Psychoanalytic Quarterly* 21.2 (1952), pp. 196–214.

[75] Felix Deutsch. “Section of psychology: Thus speaks the body - An analysis of postural behavior”. In: *Transactions of the New York Academy of Sciences* 12.2 Series II (1949), pp. 58–62. DOI: https://doi.org/10.1111/j.2164-0947.1949.tb01869.x. eprint: https://nyaspubs.onlinelibrary.wiley.com/doi/pdf/10.1111/j.2164-0947.1949.tb01869.x. URL: https://nyaspubs.onlinelibrary.wiley.com/doi/abs/10.1111/j.2164-0947.1949.tb01869.x.

[76] Jacob Devlin et al. “Bert: Pre-training of deep bidirectional transformers for language understanding”. In: *arXiv preprint arXiv:1810.04805* (2018).

[77] Anca D Dragan, Kenton CT Lee, and Siddhartha S Srinivasa. “Legibility and predictability of robot motion”. In: *2013 8th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE. 2013, pp. 301–308.

[78] Starkey Duncan. “Some signals and rules for taking speaking turns in conversations.” In: *Journal of personality and social psychology* 23.2 (1972), p. 283.

[79] Starkey Duncan and Donald W Fiske. *Face-to-face interaction: Research, methods, and theory*. Routledge, 2015.

[80] Starkey Duncan Jr and George Niederehe. “On signalling that it’s your turn to speak”. In: *Journal of experimental social psychology* 10.3 (1974), pp. 234–247.

[81] David Efron. “Gesture and environment.” In: (1941).

[82] Irenâus Eibl-Eibesfeldt. “Similarities and differences between cultures in expressive movements”. In: 1979.

[83] Irenâus Eibl-Eibesfeldt. “The expressive behavior of the deaf-and-blind born”. In: 1973.

[84] P. Ekman. “Universal and Cultural Differences in Facial Expression of Emotions”. In: ed. by J. Cole. Nebraska Symposium on Motivation. Lincoln: University of Nebraska Press, 1972, pp. 207–283.

[85] Paul Ekman. “Cross-cultural studies of facial expression”. In: *Darwin and facial expression: A century of research in review* 169222.1 (1973).

[86] Paul Ekman and Wallace V Friesen. “Facial action coding system”. In: *Environmental Psychology & Nonverbal Behavior* (1978).

[87] Paul Ekman and Wallace V Friesen. “Nonverbal leakage and clues to deception”. In: *Psychiatry* 32.1 (1969), pp. 88–106.

[88] Paul Ekman and Wallace V Friesen. “The repertoire of nonverbal behavior: Categories, origins, usage, and coding”. In: *semiotica* 1.1 (1969), pp. 49–98.

[89] Paul Ekman and Wallace V Friesen. *Unmasking the face: A guide to recognizing emotions from facial clues*. Vol. 10. Ishk, 2003.

[90] Paul Ekman, Wallace V Friesen, and Phoebe Ellsworth. *Emotion in the human face: Guidelines for research and an integration of findings*. Vol. 11. Elsevier, 2013.

[91] Paul Ekman, Wallace V. Friesen, and Phobe Ellsworth. “Emotion in the Human Face”. In: ed. by Paul Ekman, Wallace V. Friesen, and Phobe Ellsworth. Vol. 11. Pergamon General Psychology Series. Pergamon, 1972. ISBN: 978-0-08-016643-8. DOI: https://doi.org/10.1016/B978-0-08-016643-8.50006-9. URL: https://www.sciencedirect.com/science/article/pii/B9780080166438500069.

[92] Paul Ekman et al. “Kinesic cues: The body, eyes, and face.” In: (1999).

[93] Alan Elangovan. *Encyclopedia of Body Language: What Every Movement Says*. Partridge Publishing Singapore, 2020.

[94] Phoebe Ellsworth and Lee Ross. “Intimacy in response to direct gaze”. In: *Journal of experimental social psychology* 11.6 (1975), pp. 592–613.

[95] Isa N Engleberg and Dianna R Wynn. *Working in groups: Communication principles and strategies*. Pearson, 2016.

[96] Damien Erceau and Nicolas Guéguen. “Tactile contact and evaluation of the toucher”. In: *The Journal of social psychology* 147.4 (2007), pp. 441–444.

[97] Max Ernest-Jones, Daniel Nettle, and Melissa Bateson. “Effects of eye images on everyday cooperative behavior: a field experiment”. In: *Evolution and Human Behavior* 32.3 (2011), pp. 172–178.

[98] Gabriele Fanelli et al. “Random forests for real time 3d face analysis”. In: *International journal of computer vision* 101.3 (2013), pp. 437–458.

[99] Hao-Shu Fang et al. “AlphaPose: Whole-Body Regional Multi-Person Pose Estimation and Tracking in Real-Time”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2022).

[100] Hao-Shu Fang et al. “Rmpe: Regional multi-person pose estimation”. In: *Proceedings of the IEEE international conference on computer vision*. 2017, pp. 2334–2343.

[101] Julius Fast. *Body language*. Vol. 82348. Simon and Schuster, 1970.

[102] Barbara J Fehr and Ralph V Exline. “Social visual interaction: A conceptual and literature review.” In: (1987).

[103] Jeffrey D Fisher, Marvin Rytting, and Richard Heslin. “Hands touching hands: Affective and evaluative effects of an interpersonal touch”. In: *Sociometry* (1976), pp. 416–421.

[104] Kory Floyd. “All touches are not created equal: Effects of form and duration on observers’ interpretations of an embrace”. In: *Journal of Nonverbal Behavior* 23.4 (1999), pp. 283–299.

[105] Dieter Fox, Wolfram Burgard, and Sebastian Thrun. “The dynamic window approach to collision avoidance”. In: *IEEE Robotics & Automation Magazine* 4.1 (1997), pp. 23–33.

[106] Alan J. Fridlund. *Human facial expression: An evolutionary view*. Academic press, 2014.

[107] Alexandra Frischen, Andrew P Bayliss, and Steven P Tipper. “Gaze cueing of attention: visual attention, social cognition, and individual differences.” In: *Psychological bulletin* 133.4 (2007), p. 694.

[108] Tarak Gandhi and Mohan Manubhai Trivedi. “Image based estimation of pedestrian orientation for improving path prediction”. In: *2008 IEEE Intelligent Vehicles Symposium*. IEEE. 2008, pp. 506–511.

[109] Andreas Geiger et al. “Vision meets robotics: The kitti dataset”. In: *The International Journal of Robotics Research* 32.11 (2013), pp. 1231–1237.

[110] Amir Ghodrati, Marco Pedersoli, and Tinne Tuytelaars. “Is 2d information enough for viewpoint estimation?” In: *Proceedings BMVC 2014* (2014), pp. 1–12.

[111] Robert Gifford. “Personality and Nonverbal Behavior: A Complex Conundrum.” In: (2006).

[112] Maureen Gillespie et al. “Verbal working memory predicts co-speech gesture: Evidence from individual differences”. In: *Cognition* 132.2 (2014), pp. 174–180. ISSN: 0010-0277. DOI: https://doi.org/10.1016/j.cognition.2014.03.012. URL: https://www.sciencedirect.com/science/article/pii/S0010027714000511.

[113] Georgia Gkioxari et al. “Using k-poselets for detecting people and localizing their keypoints”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2014, pp. 3582–3589.

[114] Rachel Gockley, Jodi Forlizzi, and Reid Simmons. “Natural person-following behavior for social robots”. In: *Proceedings of the ACM/IEEE international conference on Human-robot interaction*. 2007, pp. 17–24.

[115] E. Goffman. *Relations in Public*. Transaction Publishers, 2009. ISBN: 9781412845199. URL: https://books.google.it/books?id=ApSW54vTsYwC.

[116] Susan Goldin-Meadow and Diane Brentari. “Gesture, sign and language: The coming of age of sign language and gesture studies”. In: *The Behavioral and brain sciences* (Oct. 2015), pp. 1–82. DOI: 10.1017/S0140525X15001247.

[117] Eberhard Graether and Florian Mueller. “Joggobot: a flying robot as jogging companion”. In: *CHI’12 Extended Abstracts on Human Factors in Computing Systems*. 2012, pp. 1063–1066.

[118] Jean Ann Graham and Michael Argyle. “A cross-cultural study of the communication of extra-verbal meaning by gestures (1)”. In: *International Journal of Psychology* 10.1 (1975), pp. 57–67.

[119] Jinwei Gu et al. “Dynamic facial analysis: From bayesian filtering to recurrent neural network”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 1548–1557.

[120] Nicolas Guéguen. “Touch, awareness of touch, and compliance with a request”. In: *Perceptual and motor skills* 95.2 (2002), pp. 355–360.

[121] Laura K Guerrero and Peter A Andersen. “The waxing and waning of relational intimacy: Touch as a function of relational stage, gender and touch avoidance”. In: *Journal of Social and Personal Relationships* 8.2 (1991), pp. 147–165.

[122] Richard F Haase and Donald T Tepper. “Nonverbal components of empathic communication.” In: *Journal of counseling psychology* 19.5 (1972), p. 417.

[123] Edward T. Hall. “A System for the Notation of Proxemic Behavior”. In: *American Anthropologist* 65.5 (1963), pp. 1003–1026. DOI: https://doi.org/10.1525/aa.1963.65.5.02a00020. eprint: https://anthrosource.onlinelibrary.wiley.com/doi/pdf/10.1525/aa.1963.65.5.02a00020. URL: https://anthrosource.onlinelibrary.wiley.com/doi/abs/10.1525/aa.1963.65.5.02a00020.

[124] Joanna Hall et al. “Perception of own and robot engagement in human–robot interactions and their dependence on robotics knowledge”. In: *Robotics and Autonomous Systems* 62.3 (2014), pp. 392–399.

[125] Judith A Hall, Erik J Coats, and Lavonia Smith LeBeau. “Nonverbal behavior and the vertical dimension of social relations: a meta-analysis.” In: *Psychological bulletin* 131.6 (2005), p. 898.

[126] Judith A. Hall. “Nonverbal sex differences : communication accuracy and expressive style”. In: 1984.

[127] Judith A. Hall and Amy G. Halberstadt. “Sex roles and nonverbal communication skills”. In: *Sex roles* 7.3 (1981), pp. 273–287.

[128] Judith A. Hall, Terrence G. Horgan, and Nora A. Murphy. “Nonverbal Communication”. In: *Annual Review of Psychology* 70.1 (2019). PMID: 30256720, pp. 271–294. DOI: 10.1146/annurev-psych-010418-103145. eprint: https://doi.org/10.1146/annurev-psych-010418-103145. URL: https://doi.org/10.1146/annurev-psych-010418-103145.

[129] T Hall Edward. *The Hidden Dimension*. 1966.

[130] Kota Hara and Rama Chellappa. “Growing regression tree forests by classification for continuous object pose estimation”. In: *International Journal of Computer Vision* 122.2 (2017), pp. 292–312.

[131] Kota Hara, Raviteja Vemulapalli, and Rama Chellappa. “Designing deep convolutional neural networks for continuous object orientation estimation”. In: *arXiv preprint arXiv:1702.01499* (2017).

[132] Kotaro Hayashi et al. “Friendly patrolling: A model of natural encounters”. In: *Proc. RSS*. 2012, p. 121.

[133] Kaiming He et al. “Deep residual learning for image recognition. CVPR. 2016”. In: *arXiv preprint arXiv:1512.03385* (2016).

[134] Marvin A. Hecht and Nalini Ambady. “Nonverbal communication and psychology: Past and future”. In: *New Jersey Journal of Communication* 7.2 (1999), pp. 156–170. DOI: 10.1080/15456879909367364. eprint: https://doi.org/10.1080/15456879909367364. URL: https://doi.org/10.1080/15456879909367364.

[135] Dan Hendrycks and Kevin Gimpel. “Gaussian error linear units (gelus)”. In: *arXiv preprint arXiv:1606.08415* (2016).

[136] Matthew J Hertenstein. “Touch: Its communicative functions in infancy”. In: *Human Development* 45.2 (2002), pp. 70–94.

[137] Matthew J Hertenstein et al. “Touch communicates distinct emotions.” In: *Emotion* 6.3 (2006), p. 528.

[138] Matthew J. Hertenstein et al. “The Communicative Functions of Touch in Humans, Nonhuman Primates, and Rats: A Review and Synthesis of the Empirical Research”. In: *Genetic, Social, and General Psychology Monographs* 132.1 (2006). PMID: 17345871, pp. 5–94. DOI: 10.3200/MONO.132.1.5-94. eprint: https://doi.org/10.3200/MONO.132.1.5-94. URL: https://doi.org/10.3200/MONO.132.1.5-94.

[139] Richard Heslin. “Steps toward a taxonomy of touching”. In: *annual convention of the Midwestern Psychological Association, Chicago*. 1974.

[140] Eckhard H Hess. “Attitude and pupil size”. In: *Scientific american* 212.4 (1965), pp. 46–55.

[141] Eckhard H Hess. “Pupillometrics. A method of studying mental, emotional, and senosory prrocesses.” In: *Handbook of psychophysiology* (1972), pp. 491–531.

[142] Eckhard H Hess and James M Polt. “Pupil size as related to interest value of visual stimuli”. In: *Science* 132.3423 (1960), pp. 349–350.

[143] Ursula Hess, Reginald B Adams, and Robert E Kleck. “Looking at you or looking elsewhere: The influence of head orientation on the signal value of emotional facial expressions”. In: *Motivation and Emotion* 31.2 (2007), pp. 137–144.

[144] Robert A Hicks and Steven Dockstader. “Cultural deprivation and pre-school children’s preferences for complex and novel stimuli”. In: *Perceptual and Motor Skills* 27.3_suppl (1968), pp. 1321–1322.

[145] M Hickson and W Self. “Biological foundations of territoriality: Nonverbal communication, language, and the law”. In: *Journal of Intercultural Communication Research* 32 (2003), pp. 265–283.

[146] Judith Holler, Kylie Turner, and Trudy Varcianna. “It’s on the tip of my fingers: Co-speech gestures during lexical retrieval in different social contexts”. In: *Language and Cognitive Processes* 28.10 (2013), pp. 1509–1518.

[147] James P. Holoka. “Nonverbal communication in the classics: research opportunities”. In: *Advances in Nonverbal Communication, Amsterdam, Benjamins* (1992), pp. 237–254.

[148] Jacob Hornik. “Tactile stimulation and consumer response”. In: *Journal of consumer research* 19.3 (1992), pp. 449–458.

[149] Jacob Hornik and Shmuel Ellis. “Strategies to secure compliance for a mall intercept interview”. In: *Public Opinion Quarterly* 52.4 (1988), pp. 539–551.

[150] Andrew G Howard et al. “Mobilenets: Efficient convolutional neural networks for mobile vision applications”. In: *arXiv preprint arXiv:1704.04861* (2017).

[151] Gao Huang et al. “Densely connected convolutional networks”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 4700–4708.

[152] Helge Hüttenrauch et al. “Investigating spatial relationships in human-robot interaction”. In: *2006 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE. 2006, pp. 5052–5059.

[153] Michita Imai, Tetsuo Ono, and Hiroshi Ishiguro. “Physical relation and expression: Joint attention for human-robot interaction”. In: *IEEE Transactions on Industrial Electronics* 50.4 (2003), pp. 636–643.

[154] Eldar Insafutdinov et al. “Deepercut: A deeper, stronger, and faster multi-person pose estimation model”. In: *European conference on computer vision*. Springer. 2016, pp. 34–50.

[155] Sergey Ioffe. “Batch renormalization: Towards reducing minibatch dependence in batch-normalized models”. In: *Advances in neural information processing systems* 30 (2017).

[156] Catalin Ionescu et al. “Human3. 6m: Large scale datasets and predictive methods for 3d human sensing in natural environments”. In: *IEEE transactions on pattern analysis and machine intelligence* 36.7 (2013), pp. 1325–1339.

[157] Hiroshi Ishiguro. “Android science”. In: *Robotics Research*. Springer, 2007, pp. 118–127.

[158] Serena Ivaldi et al. “Towards engagement models that consider individual factors in HRI: On the relation of extroversion and negative attitude towards robots to gaze and speech during a human–robot assembly task”. In: *International Journal of Social Robotics* 9.1 (2017), pp. 63–86.

[159] Michel Pierre Janisse. “PUPIL SIZE, AFFECT, AND EXPOSURE FREQUENCY.” In: *Social Behavior & Personality: an international journal* 2.2 (1974).

[160] Stanley E Jones and A Elaine Yarbrough. “A naturalistic study of the meanings of touch”. In: *Communications Monographs* 52.1 (1985), pp. 19–56.

[161] Sidney M Jourard and Jane E Rubin. “Self-disclosure and touching: A study of two modes of interpersonal encounter and their inter-relation”. In: *Journal of Humanistic Psychology* 8.1 (1968), pp. 39–48.

[162] Yusuke Kato, Takayuki Kanda, and Hiroshi Ishiguro. “May i help you?-design of human-like polite approaching behavior”. In: *2015 10th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE. 2015, pp. 35–42.

[163] Douglas Kaufman and John M Mahoney. “The effect of waitresses’ touch on alcohol consumption in dyads”. In: *The Journal of social psychology* 139.3 (1999), pp. 261–267.

[164] Francis D Kelly. “Communicational significance of therapist proxemic cues.” In: *Journal of Consulting and Clinical Psychology* 39.2 (1972), p. 345.

[165] A. KENDON. “Some uses of gesture”. In: *Perspectives on silence*. New Jersey: Ablex Publishing Corporation, 1984, pp. 215–234.

[166] Adam Kendon. “Geography of gesture”. In: *Semiotica* 37.1/2 (1981), pp. 129–163.

[167] Adam Kendon. *Gesture: Visible action as utterance*. Cambridge University Press, 2004.

[168] Adam Kendon. “The study of gesture: Some observations on its history”. In: *Semiotics 1982*. Springer, 1982, pp. 45–62.

[169] Charles E Kimble, Robert A Forte, and Joyce C Yoshikawa. “Nonverbal concomitants of enacted emotional intensity and positivity: Visual and vocal behavior 1”. In: *Journal of Personality* 49.3 (1981), pp. 271–283.

[170] Charles E Kimble and Donald A Olszewski. “Gaze and emotional expression: The effects of message positivity-negativity and emotional intensity”. In: *Journal of Research in Personality* 14.1 (1980), pp. 60–69.

[171] Chris L Kleinke. “Compliance to requests made by gazing and touching experimenters in field settings”. In: *Journal of experimental social Psychology* 13.3 (1977), pp. 218–223.

[172] Chris L Kleinke. “Gaze and eye contact: a research review.” In: *Psychological bulletin* 100.1 (1986), p. 78.

[173] Mark L Knapp. “An Historical Overview of Nonverbal Research.” In: (2006).

[174] Kheng Lee Koay et al. “Exploratory study of a robot approaching a person in the context of handing over an object.” In: *AAAI spring symposium: multidisciplinary collaboration for socially assistive robotics*. Stanford, CA. 2007, pp. 18–24.

[175] Kyveli Kompatsiari et al. “Measuring engagement elicited by eye contact in Human-Robot Interaction”. In: *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE. 2019, pp. 6979–6985.

[176] Kyveli Kompatsiari et al. “On the role of eye contact in gaze cueing”. In: *Scientific reports* 8.1 (2018), pp. 1–10.

[177] Hideki Kozima, Marek P Michalowski, and Cocoro Nakagawa. “Keepon”. In: *International Journal of social robotics* 1.1 (2009), pp. 3–18.

[178] RM Krauss, Y Chen, and RF Gottesman. “Lexical gestures and lexical access: a processing model”. In: *Language and gesture*. Ed. by McNeill D. Cambridge: Cambridge University Press., 2001, pp. 261–283.

[179] Robert M. Krauss, Yihsiu Chen, and Purnima Chawla. “Nonverbal Behavior and Nonverbal Communication: What do Conversational Hand Gestures Tell Us?” In: ed. by Mark P. Zanna. Vol. 28. Advances in Experimental Social Psychology. Academic Press, 1996, pp. 389–450. DOI: https://doi.org/10.1016/S0065-2601(08)60241-5. URL: https://www.sciencedirect.com/science/article/pii/S0065260108602415.

[180] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. “Imagenet classification with deep convolutional neural networks”. In: *Communications of the ACM* 60.6 (2017), pp. 84–90.

[181] Dana Kulić and Elizabeth A Croft. “Safe planning for human-robot interaction”. In: *Journal of Robotic Systems* 22.7 (2005), pp. 383–396.

[182] Amit Kumar, Azadeh Alavi, and Rama Chellappa. “Kepler: Keypoint and pose estimation of unconstrained faces by learning efficient h-cnn regressors”. In: *2017 12th ieee international conference on automatic face & gesture recognition (fg 2017)*. IEEE. 2017, pp. 258–265.

[183] Alina Kuznetsova et al. “The Open Images Dataset V4: Unified image classification, object detection, and visual relationship detection at scale”. In: *IJCV* (2020).

[184] Lubor Ladicky, Philip HS Torr, and Andrew Zisserman. “Human pose estimation using a joint pixel-wise and part-wise formulation”. In: *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2013, pp. 3578–3585.

[185] Sergio Lafuente-Arroyo et al. “RGB camera-based fallen person detection system embedded on a mobile platform”. In: *Expert Systems with Applications* 197 (2022), p. 116715.

[186] Divesh Lala et al. “Detection of social signals for recognizing engagement in human-robot interaction”. In: *arXiv preprint arXiv:1709.10257* (2017).

[187] Stephen RH Langton, Roger J Watt, and Vicki Bruce. “Do the eyes have it? Cues to the direction of social attention”. In: *Trends in cognitive sciences* 4.2 (2000), pp. 50–59.

[188] Donald Lateiner. “Affect displays in the epic poetry of Homer, Vergil, and Ovid”. In: *Advancements in Nonverbal Communication: Sociocultural, Clinical, Esthetic and Literary Perspectives, Poyatos, F.(coord.), Amsterdam/Filadelfia, John Benjamins* (1992), pp. 255–269.

[189] J. Laver. “The analysis of vocal quality: from the classical period to the twentieth century”. In: *Towards a history of phonetics* (1981), pp. 79–99.

[190] Andrew Lavin and Scott Gray. “Fast algorithms for convolutional neural networks”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 4013–4021.

[191] Yann LeCun et al. “Backpropagation applied to handwritten zip code recognition”. In: *Neural computation* 1.4 (1989), pp. 541–551.

[192] Hedwig Lewis. *Body language: A guide for professionals*. SAGE Publications India, 2012.

[193] Tsung-Yi Lin et al. “Microsoft coco: Common objects in context”. In: *European conference on computer vision*. Springer. 2014, pp. 740–755.

[194] Jacqueline Lindenfeld. “Verbal and non-verbal elements in discourse”. In: (1971).

[195] Stephen Littlejohn and Karen Foss. *Encyclopedia of Communication Theory*. pages 691-694. Thousand Oaks, Oct. 2009. DOI: 10.4135/9781412959384. URL: https://doi.org/10.4135/9781412959384.

[196] Peiye Liu, Wu Liu, and Huadong Ma. “Weighted sequence loss based spatial-temporal deep learning framework for human body orientation estimation”. In: *2017 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE. 2017, pp. 97–102.

[197] Wu Liu et al. “Accurate estimation of human body orientation from RGB-D sensors”. In: *IEEE Transactions on cybernetics* 43.5 (2013), pp. 1442–1452.

[198] Ze Liu et al. “Swin transformer: Hierarchical vision transformer using shifted windows”. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021, pp. 10012–10022.

[199] Zhuang Liu et al. “A ConvNet for the 2020s”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2022).

[200] Ilya Loshchilov and Frank Hutter. “Decoupled weight decay regularization”. In: *arXiv preprint arXiv:1711.05101* (2017).

[201] A Lowen. “Spiritual body”. In: *New York: Publishing Agency Jacek Santorski & CO* (1991).

[202] Alexander Lowen. *Physical Dynamics of Character Structure: Bodily Form and Movement in Analytic Therapy*. Grune & Stratton, 1971.

[203] Alexander Lowen. *The betrayal of the body*. The Alexander Lowen Foundation, 2012.

[204] Iain Matthews and Simon Baker. “Active appearance models revisited”. In: *International journal of computer vision* 60.2 (2004), pp. 135–164.

[205] Rachel I. Mayberry and Elena Nicoladis. “Gesture Reflects Language Development: Evidence From Bilingual Children”. In: *Current Directions in Psychological Science* 9.6 (2000), pp. 192–196. DOI: 10.1111/1467-8721.00092. eprint: https://doi.org/10.1111/1467-8721.00092. URL: https://doi.org/10.1111/1467-8721.00092.

[206] David McNeill. “Hand and Mind”. In: *Advances in Visual Semiotics* (1992), p. 351.

[207] Albert Mehrabian. “Inference of attitudes from the posture, orientation, and distance of a communicator.” In: *Journal of consulting and clinical psychology* 32.3 (1968), p. 296.

[208] Spero A Metalis, Eckhard H Hess, and Paul W Beaver. “Pupillometric analysis of two theories of obesity”. In: *Perceptual and Motor Skills* 55.1 (1982), pp. 87–92.

[209] Sushmita Mitra and Tinku Acharya. “Gesture recognition: A survey”. In: *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)* 37.3 (2007), pp. 311–324.

[210] mon95. *Mon95/sign-language-and-static-gesture-recognition-using-sklearn: A machine learning pipeline that performs hand localization and static-gesture recognition built using the scikit learn and scikit image libraries*. URL: https://github.com/mon95/Sign-Language-and-Static-gesture-recognition-using-sklearn.

[211] Nina-Jo Moore, M Hickson, and DW Stacks. *Nonverbal communication*. New York: Oxford University Press, 2010.

[212] Luis Yoichi Morales Saiki et al. “How do people walk side-by-side? Using a computational model of human behavior for a social robot”. In: *Proceedings of the seventh annual ACM/IEEE international conference on Human-Robot Interaction*. 2012, pp. 301–308.

[213] Hans Moravec. *Mind children: The future of robot and human intelligence*. Harvard University Press, 1988.

[214] Desmond Morris et al. *Gestures, their origins and distribution*. Stein & Day Pub, 1979.

[215] Emily Mower et al. "Investigating implicit cues for user state estimation in human-robot interaction using physiological measurements". In: *RO-MAN 2007-The 16th IEEE International Symposium on Robot and Human Interactive Communication*. IEEE. 2007, pp. 1125–1130.

[216] Bilge Mutlu et al. "Conversational gaze mechanisms for humanlike robots". In: *ACM Transactions on Interactive Intelligent Systems (TiiS)* 1.2 (2012), pp. 1–33.

[217] Bilge Mutlu et al. "Footing in human-robot conversations: how robots might shape participant roles using gaze cues". In: *Proceedings of the 4th ACM/IEEE international conference on Human robot interaction*. 2009, pp. 61–68.

[218] Vinod Nair and Geoffrey E Hinton. "Rectified linear units improve restricted boltzmann machines". In: *Icml*. 2010.

[219] Chikahito Nakajima et al. "Full-body person recognition system". In: *Pattern recognition* 36.9 (2003), pp. 1997–2006.

[220] Yasushi Nakauchi and Reid Simmons. "A social robot that stands in line". In: *Autonomous Robots* 12.3 (2002), pp. 313–324.

[221] Jane C Nannberg and Christine H Hansen. "Post-compliance touch: An incentive for task performance". In: *The Journal of Social Psychology* 134.3 (1994), pp. 301–307.

[222] Alejandro Newell, Kaiyu Yang, and Jia Deng. "Stacked hourglass networks for human pose estimation". In: *European conference on computer vision*. Springer. 2016, pp. 483–499.

[223] Tatsuya Nomura et al. "Prediction of human behavior in human–robot interaction using psychological scales for anxiety and negative attitudes toward robots". In: *IEEE transactions on robotics* 24.2 (2008), pp. 442–451.

[224] Gabriel L Oliveira et al. "Deep learning for human part discovery in images". In: *2016 IEEE International conference on robotics and automation (ICRA)*. IEEE. 2016, pp. 1634–1641.

[225] Suzanne Oosterwijk et al. "Embodied emotion concepts: How generating words about pride and disappointment influences posture". In: *European Journal of Social Psychology* 39.3 (2009), pp. 457–466. DOI: https://doi.org/10.1002/ejsp.584. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/ejsp.584. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/ejsp.584.

[226] Harriet Oster and Paul Ekman. "Facial behavior in child development". In: *Minnesota symposia on child psychology*. Vol. 11. Erlbaum Hillsdale, NJ. 1978, pp. 231–276.

[227] Massimiliano Patacchiola and Angelo Cangelosi. “Head pose estimation in the wild using convolutional neural networks and adaptive gradient methods”. In: *Pattern Recognition* 71 (2017), pp. 132–143.

[228] Miles L Patterson, Jack L Powell, and Mary G Lenihan. “Touch, compliance, and interpersonal affect”. In: *Journal of Nonverbal behavior* 10.1 (1986), pp. 41–50.

[229] Allan Pease and Barbara Pease. *The Definitive Book of Body Language*. Pease International, 2004.

[230] David I Perrett and Nathan J Emery. “Understanding the intentions of others from visual signals: neurophysiological evidence.” In: (1994).

[231] Bull Peter. *Body movement and interpersonal communication*. Chichester ; New York : Wiley, 1983, pp. 79–87.

[232] Leonid Pishchulin et al. “Articulated people detection and pose estimation: Reshaping the future”. In: *2012 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE. 2012, pp. 3178–3185.

[233] Leonid Pishchulin et al. “Deepcut: Joint subset partition and labeling for multi person pose estimation”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 4929–4937.

[234] Robert E Pittenger, Charles F Hockett, and John J Danehy. “The first five minutes: A sample of microscopic interview analysis.” In: (1960).

[235] James M Polt and Eckhard H Hess. “Changes in pupil size to visually presented words”. In: *Psychonomic Science* 12.8 (1968), pp. 389–390.

[236] Jordi Pont-Tuset et al. “Connecting Vision and Language with Localized Narratives”. In: *ECCV*. 2020.

[237] *Posture*. `https://www.physio-pedia.com/Posture#cite_note-1`.

[238] P Pramod Kumar, Prahlad Vadakkepat, and Loh Ai Poh. “The NUS hand posture datasets I”. In: (2017).

[239] P Pramod Kumar, Prahlad Vadakkepat, and Loh Ai Poh. “The NUS hand posture datasets II”. In: (2017).

[240] Rosenthal R. et al. *Sensitivity to non-verbal communication: the PONS Test*. Johns Hopkins University Press, Baltimore, 1979.

[241] Alec Radford et al. “Language models are unsupervised multitask learners”. In: *OpenAI blog* 1.8 (2019), p. 9.

[242] Ilija Radosavovic et al. “Designing network design spaces”. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2020, pp. 10428–10436.

[243] Ilija Radosavovic et al. “On network design spaces for visual recognition”. In: *Proceedings of the IEEE/CVF international conference on computer vision*. 2019, pp. 1882–1890.

[244] Prajit Ramachandran et al. “Stand-alone self-attention in vision models”. In: *Advances in Neural Information Processing Systems* 32 (2019).

[245] Rahul Raman et al. “Direction estimation for pedestrian monitoring system in smart cities: An HMM based approach”. In: *IEEE Access* 4 (2016), pp. 5788–5808.

[246] Rajeev Ranjan, Vishal M Patel, and Rama Chellappa. “Hyperface: A deep multi-task learning framework for face detection, landmark localization, pose estimation, and gender recognition”. In: *IEEE transactions on pattern analysis and machine intelligence* 41.1 (2017), pp. 121–135.

[247] Rajeev Ranjan et al. “An all-in-one convolutional neural network for face analysis”. In: *2017 12th IEEE international conference on automatic face & gesture recognition (FG 2017)*. IEEE. 2017, pp. 17–24.

[248] Mudassar Raza et al. “Appearance based pedestrians’ head pose and body orientation estimation using deep learning”. In: *Neurocomputing* 272 (2018), pp. 647–659.

[249] William K Redican. “An evolutionary perspective on human facial displays”. In: *Emotion in the human face* 2 (1982), pp. 212–280.

[250] Clare S Rees et al. “Back and neck pain are related to mental health problems in adolescence”. In: *BMC public health* 11.1 (2011), pp. 1–8.

[251] Wilhelm Reich. “On character analysis”. In: *The Psychoanalytic Review (1913-1957)* 20 (1933), p. 89.

[252] Shaoqing Ren et al. “Faster r-cnn: Towards real-time object detection with region proposal networks”. In: *Advances in neural information processing systems* 28 (2015).

[253] Charles Rich et al. “Recognizing engagement in human-robot interaction”. In: *2010 5th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE. 2010, pp. 375–382.

[254] Laurel D Riek, Philip C Paul, and Peter Robinson. “When my robot smiles at me: Enabling human-robot rapport via real-time head gesture mimicry”. In: *Journal on Multimodal User Interfaces* 3.1 (2010), pp. 99–108.

[255] Margaret Gwendoline Riseborough. “Physiographic gestures as decoding facilitators: Three experiments exploring a neglected facet of communication”. In: *Journal of Nonverbal Behavior* 5.3 (1981), pp. 172–183.

[256] John H. Riskind and Carolyn C. Gotay. “Physical posture: Could it have regulatory or feedback effects on motivation and emotion?” In: *Motivation and Emotion* 6.3 (Sept. 1982), pp. 273–298. ISSN: 1573-6644. DOI: 10.1007/BF00992249. URL: https://doi.org/10.1007/BF00992249.

[257] Ben Robins et al. “Robot-mediated joint attention in children with autism: A case study in robot-human interaction”. In: *Interaction studies* 5.2 (2004), pp. 161–198.

[258] Probabilistic Robotics. *Sebastian Thrun, Wolfram Burgard and Dieter Fox*. 2005.

[259] William T. Rogers. “The contribution of kinesic illustrators toward the comprehension of verbal behavior within utterances”. In: *Human communication research* 5.1 (1978), pp. 54–62.

[260] John W Rohrbaugh. “The orienting reflex: Performance and central nervous system manifestations”. In: *Varieties of attention* (1984), pp. 323–373.

[261] Irene Rossberg-Gempton and Gary D. Poole. “The effect of open and closed postures on pleasant and unpleasant emotions”. In: *The Arts in Psychotherapy* 20.1 (1993). Special Issue Research in the Creative Arts Therapies, pp. 75–82. ISSN: 0197-4556. DOI: https://doi.org/10.1016/0197-4556(93)90034-Y. URL: https://www.sciencedirect.com/science/article/pii/019745569390034Y.

[262] Vincent Rousseau et al. “Sorry to interrupt, but may I have your attention? Preliminary design and evaluation of autonomous engagement in HRI”. In: *Journal of Human-Robot Interaction* 2.3 (2013), pp. 41–61.

[263] Nataniel Ruiz, Eunji Chong, and James M. Rehg. “Fine-Grained Head Pose Estimation Without Keypoints”. In: *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*. June 2018.

[264] Olga Russakovsky et al. “Imagenet large scale visual recognition challenge”. In: *International journal of computer vision* 115.3 (2015), pp. 211–252.

[265] Derek R Rutter. *Looking and seeing: The role of visual communication in social interaction*. Wiley, 1984.

[266] Daisuke Sakamoto et al. “Android as a telecommunication medium with a human-like presence”. In: *2007 2nd ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE. 2007, pp. 193–200.

[267] Hanan Salam and Mohamed Chetouani. “A multi-level context-based modeling of engagement in human-robot interaction”. In: *2015 11th IEEE international conference and workshops on automatic face and gesture recognition (FG)*. Vol. 3. IEEE. 2015, pp. 1–6.

[268] Maha Salem et al. “To err is human (-like): Effects of robot gesture on perceived anthropomorphism and likability”. In: *International Journal of Social Robotics* 5.3 (2013), pp. 313–323.

[269] Tim Salimans and Durk P Kingma. “Weight normalization: A simple reparameterization to accelerate training of deep neural networks”. In: *Advances in neural information processing systems* 29 (2016).

[270] David Sander et al. “Interaction effects of perceived gaze direction and dynamic facial expression: Evidence for appraisal theories of emotion”. In: *European Journal of Cognitive Psychology* 19.3 (2007), pp. 470–480.

[271] Jyotirmay Sanghvi et al. “Automatic analysis of affective postures and body motion to detect engagement with a game companion”. In: *Proceedings of the 6th international conference on Human-robot interaction*. 2011, pp. 305–312.

[272] Benjamin Sapp, Alexander Toshev, and Ben Taskar. “Cascaded models for articulated pose estimation”. In: *European conference on computer vision*. Springer. 2010, pp. 406–420.

[273] Jason M Saragih, Simon Lucey, and Jeffrey F Cohn. “Deformable model fitting by regularized landmark mean-shift”. In: *International journal of computer vision* 91.2 (2011), pp. 200–215.

[274] Satoru Satake et al. “How to approach humans? Strategies for social robots to initiate interaction”. In: *Proceedings of the 4th ACM/IEEE international conference on Human robot interaction*. 2009, pp. 109–116.

[275] Brian Scassellati. “Imitation and mechanisms of joint attention: A developmental structure for building social skills on a humanoid robot”. In: *International Workshop on Computation for Metaphors, Analogy, and Agents*. Springer. 1998, pp. 176–195.

[276] Albert E Scheflen. *Communicational structure: Analysis of a psychotherapy transaction.* Indiana U. Press, 1973.

[277] Albert E Scheflen. “The significance of posture in communication systems”. In: *Psychiatry* 27.4 (1964), pp. 316–331.

[278] Leonhard Schilbach et al. “Minds made for sharing: initiating joint attention recruits reward-related neurocircuitry”. In: *Journal of cognitive neuroscience* 22.12 (2010), pp. 2702–2715.

[279] Jean-Claude Schmitt. “The rational of gestures in the west: A history from the 3rd to the 13th centuries”. In: *Advances in Nonverbal Communication: Sociocultural, Clinical, Esthetic, and Literary Perspectives. Ferdinand Poyatos, ed* (1992), pp. 77–95.

[280] Peter J. Schulz and Paul Cobley. *Handbooks of Communication Science*. Berlin, Germany: De Gruyter Mouton, 2013.

[281] Pierre Sermanet et al. “Overfeat: Integrated recognition, localization and detection using convolutional networks”. In: *arXiv preprint arXiv:1312.6229* (2013).

[282] Claude Elwood Shannon. “A mathematical theory of communication”. In: *The Bell system technical journal* 27.3 (1948), pp. 379–423.

[283] Hiroaki Shimizu and Tomaso Poggio. “Direction estimation of pedestrian from multiple still images”. In: *IEEE Intelligent Vehicles Symposium, 2004*. IEEE. 2004, pp. 596–600.

[284] Candace L Sidner and Myrosia Dzikovska. “Human-robot interaction: Engagement between humans and robots for hosting activities”. In: *Proceedings. Fourth IEEE International Conference on Multimodal Interfaces*. IEEE. 2002, pp. 123–128.

[285] Candace L Sidner et al. “Explorations in engagement for humans and robots”. In: *Artificial Intelligence* 166.1-2 (2005), pp. 140–164.

[286] Candace L Sidner et al. “Where to look: a study of human-robot engagement”. In: *Proceedings of the 9th international conference on Intelligent user interfaces*. 2004, pp. 78–84.

[287] Karen Simonyan and Andrew Zisserman. “Two-stream convolutional networks for action recognition in videos”. In: *Advances in neural information processing systems* 27 (2014).

[288] Karen Simonyan and Andrew Zisserman. “Very deep convolutional networks for large-scale image recognition”. In: *arXiv preprint arXiv:1409.1556* (2014).

[289] Ashish Singh and James E Young. “Animal-inspired human-robot interaction: A robotic tail for communicating state”. In: *2012 7th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE. 2012, pp. 237–238.

[290] Emrah Akin Sisbot et al. “A human aware mobile robot motion planner”. In: *IEEE Transactions on Robotics* 23.5 (2007), pp. 874–883.

[291] David E Smith, Joseph A Gier, and Frank N Willis. “Interpersonal touch and compliance with a marketing request”. In: *Basic and Applied Social Psychology* 3.1 (1982), pp. 35–38.

[292] Agnieszka Sorokowska et al. “Preferred Interpersonal Distances: A Global Comparison”. In: *Journal of Cross-Cultural Psychology* 48.4 (2017), pp. 577–592. DOI: 10.1177/0022022117698039. eprint: https://doi.org/10.1177/0022022117698039. URL: https://doi.org/10.1177/0022022117698039.

[293] Ke Sun et al. “Deep high-resolution representation learning for human pose estimation”. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2019, pp. 5693–5703.

[294] Min Sun, Pushmeet Kohli, and Jamie Shotton. “Conditional regression forests for human pose estimation”. In: *2012 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE. 2012, pp. 3394–3401.

[295] Daniel Szafir, Bilge Mutlu, and Terry Fong. “Communicating directionality in flying robots”. In: *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction*. 2015, pp. 19–26.

[296] Christian Szegedy et al. “Going deeper with convolutions”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015, pp. 1–9.

[297] Christian Szegedy et al. “Rethinking the inception architecture for computer vision”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 2818–2826.

[298] Marcell Székely and John Michael. “Investing in commitment: Persistence in a joint action is enhanced by the perception of a partner’s effort”. In: *Cognition* 174 (2018), pp. 37–42.

[299] Mingxing Tan and Quoc Le. “Efficientnet: Rethinking model scaling for convolutional neural networks”. In: *International conference on machine learning*. PMLR. 2019, pp. 6105–6114.

[300] Tecperson. *Sign language mnist*. Oct. 2017. URL: `https://www.kaggle.com/datasets/datamunge/sign-language-mnist`.

[301] Hugo Touvron et al. “Training data-efficient image transformers & distillation through attention”. In: *International Conference on Machine Learning*. PMLR. 2021, pp. 10347–10357.

[302] J Gregory Trafton et al. “Enabling effective human-robot interaction using perspective-taking in robots”. In: *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans* 35.4 (2005), pp. 460–470.

[303] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. “Instance normalization: The missing ingredient for fast stylization”. In: *arXiv preprint arXiv:1607.08022* (2016).

[304] Ashish Vaswani et al. “Attention is all you need”. In: *Advances in neural information processing systems* 30 (2017).

[305] Michael L Walters et al. “The influence of subjects’ personality traits on personal spatial zones in a human-robot interaction experiment”. In: *ROMAN 2005. IEEE International Workshop on Robot and Human Interactive Communication, 2005.* IEEE. 2005, pp. 347–352.

[306] Joseph B. Walther. “Nonverbal dynamics in computer-mediated communication, or:(and the net:(‘s with you:) and you:) alone”. In: *Handbook of nonverbal communication* (2006), pp. 461–479.

[307] Paul Watzlawick, Janet Beavin Bavelas, and Don D Jackson. *Pragmatics of human communication: A study of interactional patterns, pathologies and paradoxes*. WW Norton & Company, 2011.

[308] Shih-En Wei et al. “Convolutional pose machines”. In: *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*. 2016, pp. 4724–4732.

[309] Morton Wiener et al. “Nonverbal behavior and nonverbal communication.” In: *Psychological review* 79.3 (1972), p. 185.

[310] Ross Wightman, Hugo Touvron, and Hervé Jégou. “Resnet strikes back: An improved training procedure in timm”. In: *arXiv preprint arXiv:2110.00476* (2021).

[311] Frank N Willis and Helen K Hamm. “The use of interpersonal touch in securing compliance”. In: *Journal of Nonverbal Behavior* 5.1 (1980), pp. 49–55.

[312] Paul Wills et al. “Socially contingent humanoid robot head behaviour results in increased charity donations”. In: *2016 11th ACM/IEEE international conference on human-robot interaction (HRI)*. IEEE. 2016, pp. 533–534.

[313] Chenyan Wu et al. “MEBOW: Monocular Estimation of Body Orientation In the Wild”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2020, pp. 3451–3461.

[314] Yuxin Wu and Kaiming He. “Group normalization”. In: *Proceedings of the European conference on computer vision (ECCV)*. 2018, pp. 3–19.

[315] Yuxin Wu and Justin Johnson. “Rethinking” batch” in batchnorm”. In: *arXiv preprint arXiv:2105.07576* (2021).

[316] Agnieszka Wykowska et al. “Beliefs about the minds of others influence how we process sensory information”. In: *PloS one* 9.4 (2014), e94339.

[317] Agnieszka Wykowska et al. “Humans are well tuned to detecting agents among non-agents: examining the sensitivity of human perception to behavioral characteristics of intentional systems”. In: *International Journal of Social Robotics* 7.5 (2015), pp. 767–781.

[318] Saining Xie et al. “Aggregated residual transformations for deep neural networks”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 1492–1500.

[319] Cai Xing and Derek M Isaacowitz. “Aiming at happiness: How motivation affects attention to and memory for emotional images”. In: *Motivation and emotion* 30.3 (2006), pp. 243–250.

[320] Junchao Xu et al. “Robot mood is contagious: effects of robot body language in the imitation game”. In: *AAMAS*. 2014, pp. 973–980.

[321] Fumitaka Yamaoka et al. “” Lifelike” behavior of communication robots based on developmental psychology findings”. In: *5th IEEE-RAS International Conference on Humanoid Robots, 2005.* IEEE. 2005, pp. 406–411.

[322] Heng Yang et al. “Face alignment assisted by head pose estimation”. In: *arXiv preprint arXiv:1507.03148* (2015).

[323] Bangpeng Yao and Li Fei-Fei. “Recognizing human-object interactions in still images by modeling the mutual context of objects and human poses”. In: *IEEE transactions on pattern analysis and machine intelligence* 34.9 (2012), pp. 1691–1703.

[324] Dameng Yu et al. “Continuous pedestrian orientation estimation using human keypoints”. In: *2019 IEEE International Symposium on Circuits and Systems (ISCAS)*. IEEE. 2019, pp. 1–5.

[325] Sangdoo Yun et al. “Cutmix: Regularization strategy to train strong classifiers with localizable features”. In: *Proceedings of the IEEE/CVF international conference on computer vision*. 2019, pp. 6023–6032.

[326] Leslie A. Zebrowitz. *Reading faces: Window to the soul?* Routledge, 2018.

[327] Hongyi Zhang et al. “mixup: Beyond empirical risk minimization”. In: *arXiv preprint arXiv:1710.09412* (2017).

[328] Guangzhe Zhao et al. “Video based estimation of pedestrian walking direction for pedestrian protection system”. In: *Journal of Electronics (China)* 29.1 (2012), pp. 72–81.

[329] Zhun Zhong et al. “Random erasing data augmentation”. In: *Proceedings of the AAAI conference on artificial intelligence*. Vol. 34. 07. 2020, pp. 13001–13008.

[330] Xiangxin Zhu and Deva Ramanan. “Face detection, pose estimation, and landmark localization in the wild”. In: *2012 IEEE conference on computer vision and pattern recognition*. IEEE. 2012, pp. 2879–2886.

[331] Xiangyu Zhu et al. “Face alignment across large poses: A 3d solution”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 146–155.