Master of Science in Informatics at Grenoble (MoSIG)
Master Mathématiques Informatique - spécialité Informatique
option Graphics, Vision and Robotics (G.V.R.)

Detection of non-verbal communication cues using multi-modal sensors : engagement detection

Wafa BENKAOUAR

PRIMA Team, INRIA

Under the supervision of Dr. Dominique VAUFREYDAZ, PRIMA - INRIA, Université Pierre-Mendès-France

Defended before a jury composed of:

Prof. James CROWLEY
Dr. Rémi RONFARD
Prof. Noël DE-PALMA
Dr. Michel VACHER

June 2012

Acknowledgments

I am very grateful to my supervisor Dr Dominique Vaufreydaz, for his support, his guidance along this project and for our numerous discussions from which I have learned a lot.

I would like to thank also the PRIMA members, for their advices and help reviewing this report. I also thanks them for integrating me so well in the team and for making me improve my UT score over the time of this project.

Résumé

La reconnaissance d'intentions est un processus cognitif que tout un chacun effectue constamment sans en avoir conscience. Cette capacité permet une anticipation qui rend les échanges interactifs plus fluides. Dans le cadre de l'intention d'interagir, des signaux non verbaux sont utiliser afin de communiquer cette attention à l'interlocuteur. Ce projet vise à détecter ces signaux afin de permettre à un robot de savoir quand il va être solliciter pour une interaction.

Classiquement la distance humain-robot sert à la détection d'engagement, mais cette méthode n'est pertinente que dans le cadre d'un interface immobile. Dans le cadre d'un robot compagnon, ce critère spatial ne peut être suffisant pour la détection d'intention d'interaction. En effet, le robot est capable de se déplacer dans l'appartement. Cette détection d'engagement est améliorée par notre approche intégrant des caractéristiques multimodales au lieu de seulement prendre une décision en fonction de la distance humain-robot. L'utilisation du capteur Kinect de Microsoft [1] vise à augmenter la réutilisabilité d'un tel détecteur à tout robot équipé.

Le premier défit a été la construction d'un corpus de données multimodales, l' étiquettage temporel d' évenemment de chacune de ces modalité. Cet acquisition simultanée de données á partir de neuf canaux de communication en parallèle a exigé une optimisation de l'acquisition de tous les capteurs afin de maintenir une fréquence d'enregistrement qui convient à la détection d'intention. La fusion de données hétérogènes pour la prise de décision a requis une synchronicité de celles-ci. La validation expérimentale montre que la multimodalité apporte plus de précision que le détecteur basé uniquement sur les données spatiales. La confrontation de cette méthode de détection à des données collectés dans des conditions spontanées souligne sa robustesse et permet de la déployée en environnement réel.

Abstract

Recognition of intentions is an unconscious cognitive process vital to human communication. This skill enables anticipation and this aides interactive exchange between humans. Within the context of intention for interaction, non-verbal signals are used to communicate this intention to the partner. In this research project we have investigated methods to detect these signals in order to permit a robot to know when it is about to be addressed.

Classically, the human-robot distance is used to detect the engagement. However this method is relevant only for a static interactive interface. The approach chosen to improve such detection is to integrate multi-modal characteristics instead of only using the distance human-robot to take the decision. The use of the Kinect sensors attempt to have a better re-usability of such a detector to any robot equipped.

Our first challenge has been to build a multi-modal corpus of data with a temporal and event labelling for each of the modalities Simultaneous recording of nine channels of communication in parallel required an optimization of the acquisition from a suit of sensors in order to maintain the necessary acquisition frequency. Synchronization of the heterogeneous data was required for fusion in order to reach a decision. Experimental validation shows that the use of multi-modal sensors gives more precision to the detector than the unique use of spatial features. The evaluation of this new method of detection to data collected in spontaneous conditions highlight its robustness and validates use of such technique in real environment.

Contents

List of Figures

1 Introduction

1.1 Context of Multi-modal Non-Verbal Communication

A conversation between two individuals is a form of interaction. This interaction is characterized by the social signals sent and interpreted by each persons as well as their reactions to signals perceived from their partner. In other words, interaction is a reciprocal exchange of signals that affects the participants involved in the interaction. The most common example of interaction is communication.

Social signal processing and affective computing have emerged as new areas of Computer Sciences over the last ten years (R. Picard [6]). These new areas explore technique for the multi-modal aspect of the human communication in order to develop more natural interaction between humans and robots. Speech is an important channel for communication and speech recognition requires audio and signal processing as well as semantics and linguistics domain. In addition to the semantic of the speech, other channels convey emotions, and the inner goals of humans. For this reason, the research community is increasingly interested in non-verbal communication. Whereas, speech is mainly mono-modal, non-verbal communication (NVC) uses a variety of channels to convey messages.

Paralinguistic signals refer to the signals accompanying speech but separated from the actual language. Examples of such include the tone of the voice, the loudness and the variations of pitch. These signals affect the semantic of the speech itself. Paralinguistic signals are included within a large class, the Non-Verbal signals. Where paralinguistic non-verbal signals are present only when we speak, non-verbal signals are more largely related to all the signals conveyed by other channels than spoken language. These signals can be conscious or not, and often reflect our intentions and emotions.

Beside the wish to detect our emotions, research in multi-modal non-verbal communication is also motivated by making the human-machine interaction more natural by allowing the user to use intuitive communication modalities instead of the mouse and the keyboard. The gesture and audio control have a large field of applications, especially in assistance for handicapped persons and elders.

Non-verbal communications also impact research in Robotics and Human-Robots Interaction. In the context of human-robot interaction, companion robots should also be able to detect the intentions of humans in order to adapt their behaviour during dialog with humans. Intention recognition allows the interacting agent to take quick decisions and to respond better to the user's need.

The recognition of the intention is one of the new challenges of robotics. Indeed, for natural interaction human-robot, the intention reading of the behavioural cues from an individual is fundamental. Recognition of intention is a basic skill acquired by infants early in their development. Vernon in [7] states that one of among other skills; the perception of the direction of the attention of others is crucial for the infant to master social interactions. The perception of intentions and emotions, present in newborn infants, helps to set their "preparedness" for social interaction.

In neurocognition, the Broca's area responsible of language comprehension, action recognition and prediction, and speech-associated gestures would be the host of intention recognition in the human brain. According to Vernon, studies have shown that the activation of the Broca's area is significantly higher when a subject observes goal-directed actions with intentional cues rather than meaningless gestures. Human cognition has a high part of anticipation, allowing reading the intentions, and guessing the goal in order to react quickly to some stimulus. This skill is very important for turn-taking in interaction.

While recognition of human action is an active area of research, there are no established techniques for recognition of intention of a human that has been modelled, even if this prediction could be useful for a companion robot in order to take quicker decision and to be more adaptive to the human needs.

1.2 Motivation Behind the Engagement Detection

The goal of this project is to investigate techniques to detect and recognize signals for non-verbal communication reflecting the intentions and in particular the engagement of a human with a robot. We define engagement as the phase during which the human expresses the intention of an interaction. Engagement of interaction refers to the process by which two or more participants perceive, establish or maintain a dialog. Perception of engagement refers to the perception of the intention for interaction. Engagement is a real question especially when it comes to environments such as the work place or home; where people are not used to interacting with robots as shown in [8]. Conversational engagement is fundamental for communication between human users and interactive robots.

In order to learn and recognize the intention elicited by non-verbal communication acts, propositions for use of multi-modal sensors have been made. In the context of a robot companion interaction, in this project we focus on detecting the engagement of a user with a Kompai robot equipped by a Kinect device as presented in Chapter 3.

Classically, the criteria for a user's engagement is the spatial distance between the user and the communicant interface [9] . Some investigations have improved on this idea by also considering the speed of movement of the user [10]. These studies have chosen to use spatial position as criteria, and a simple assumption has been made: if the user is close to the robot, he wants to interact. This detector of engagement based on distance and sometimes speed of the human gives good

results for kiosk interface and immobile interface, yet, for an assistant living robot in real-life, close distance does not signal a desire for engagement. Indeed, many times during the day one can pass in front of the refrigerator without the wish to open it. In the same vein, the robot in order to have more human acceptable behaviour should be able to detect when itÕs about to be solicited, and to anticipate the interaction.

Considering that the robot is able to move in space, spatial features are not sufficient in home environment to characterize engagement. Indeed, in the context of a companion robot the constant proximity of the robot with a person should be a continuous trigger for interaction. Other criteria can be taken into account such as the posture, the sound and other features described below.

1.3 Proposition of Multi-modal Engagement Detection

In this study we propose a multi-modal approach for detecting engagement using the Kinect sensors [1]. The use of the Kinect sensors aims to improve the reusability of such detector, and enables us to build a detector deployable in real-life situations.

The Kinect Sensor usage has been growing since its release. The latest version of the sensor is finding many new applications, on the desktop or embedded as a sensor on robot companions. One of the advantages of the Kinect sensor is to have multiple sensors integrated in the same device. The features offered by these sensors such as the depth, the skeleton tracking, the camera and the microphones array make it a rich tool. This work intends to use these Kinect features in social signal processing for engagement detector. This prospective project aims to build a vector compose of multi-modal features useful for the description, recognition and discrimination of the engagement event by a robot companion equipped with a Kinect device.

From the literature, in particular from the cognitive sciences literature, we found some cues to measure the engagement of a person into an interaction. Hence, we propose to take into account the spatial information, body pose, frontal face detection, speech detection and sound localization in order to model the engagement detection system.

An important contribution of this work is the multi-modal dataset establish from the robot point of view. This dataset offers a realistic framework to test our hypothesis. Optimizations of the acquisition of the data were needed to limit the loss of information and to facilitate the synchronization of the multi-modal data.

1.4 Results and Impacts

Multi-class Support Vector Machine technique employed to classify the features computed form the dataset have given better results in the multi-modal condition when compared to a mono-modal spatial condition. Indeed, this work shows first that the range data used for engagement detection as spatial and speed features are not enough in a home environment. Secondly, the multi-modality gives significantly better results in the detection of engagement than the mono-modality.

With more and more powerful embedded system on the robots, we can expect such multi-modal detection to be used in real-time and to allow robot to predict the intention of interaction. Indeed, with an enriched corpus of data, a model can be trained to include temporal dimension into this system. The prediction of the engagement is a first step toward a smoother interaction human-robot.

1.5 Report Overview

Chapter 2 presents a state of the art of the domains related to this matter. This state of the art deals with non-verbal communication, multi-modality detection, fusion for engagement and detection of the humansÕ interaction intention with the robot companion. A multi-modal dataset was made due to its absence for detection of engagement. Chapter 3 explains the way the dataset has been built, the context, hardware constraints and the scenarios involved. Chapter 4 described the features extracted from these data, the synchronization and labelling, and the choices in term of classifications. Chapter 5 shows the results are using real data. Finally, the conclusion, chapter 6 summarize the accomplished work. We then evaluate the impact and the further directions opened by this project.

2 Multi-modal Social Signal Processing For Non-Verbal Communication

During a conversation, social signals are exchanged by the protagonists. The speaking person delivers a speech signal intentionally, which allows the listener to understand the intended message. Over the years, computer science has made a lot of improvement in semantic interpretation of the speech. Nevertheless, while we are having a discussion with a person, all our senses are open, and we use them to interpret, understand and react to situations. The new field of social signal processing, for the past few years, has interested more and more research in human machine interaction, cognitive sciences and robotic sciences.

In this chapter, the context of this work is presented. First, in section 2.1, the social signal processing goals are introduced. Then, non-verbal communication (NVC) is briefly defined in section 2.2. Some details about the information conveyed by NVC is given in 2.2.1, and the specific subfield of detection of engagement will be presented 2.2.2. Non-invasive modalities used for social signal processing focusing in detection of engagement and attitudes are described in **??**. Finally, as the recognition of cues of engagement is closely related to the data available, section 2.4 explains what kind of multi-modal data are currently available, how multi-modal data fusion are acquired, what are the methods of fusion and their limitations.

Communication definition In the first model of communication stated by Shannon and Weaver in the 1950 's, the message sent by a transmitter is encoded and the receiver has to decode this message in order to get the meaning of it. The transmitter and the receiver have to share the same cryptologic model. In the case where we have a verbal signal (i.e. a speech), and as soon as we know the language (i.e. the encoding algorithm used by our interlocutor), we can understand and interpret this signal. The Shannon-Weaver model is linear as described in [11] and therefore is not taking into account the reactivity in the communication. Several models succeeded integrating the reaction of the listener and building non-linear models such as Dance's Helical Spiral Model of communication. Nevertheless, the first person integrating the context and non-verbal communication cues in a communication model was Barlund in 1970 with his Transactional Model of communication. Barlund integrates explicitly in his model three signs or cues that can elicit sense or a meaning: public cues (from the environment), private object of orientation (any kind of sensory cues or object in the environment), and finally behavioural and

non-verbal cues. Hence a communicant agent does not use the only verbal channel but many channels to send and receive various messages during the speech. Human communication is hence multi-modal. In order to make human-robot communication fluent, the robot has to be able to decode these behavioural and non-verbal cues.

2.1 Social Signal Processing

According to the Social Signal Processing Network (SSPNet) in [12] social signals give information about emotions and social relationship of individuals. The SSPnet research agenda [13] gives some nomenclature of social signal processing and some guidance of research for the next few years. Humans are doted by range of abilities called social intelligence. These abilities include the ability to express and recognize social signals produced during social interactions like agreement, politeness, empathy, friendliness, conflict, etc.., coupled with the ability to manage them in order to get along well with others while winning their cooperation. Social Signal Processing (SSP) is the new research domain that aims at understanding and modelling social interactions (cognitive sciences), and at providing computers with similar abilities in human-computer interaction scenarios (human machine interaction).

As pictured in the Figure 2.1, multiple behavioural cues (voice, posture, gaze, interpersonal distance, gestures etc.) can be used as a social signal to express an emotion, or a mood (in this case aggressive or disagreement). Even with the simple silhouettes of the individuals in interaction, we can relatively guess their mood.

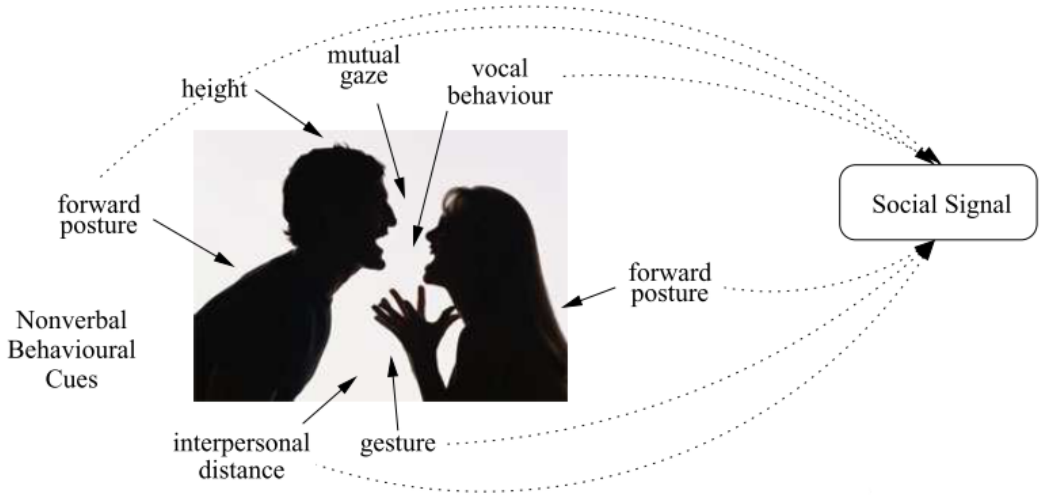


Figure 2.1: Behavioural cues and social signals, in [2]

An intelligent agent is commonly defined as an agent who is to perceive, to learn, and to adapt to the world. Social signals are manifested through a multiplicity of

non-verbal behavioural cues including facial expressions, body postures & gestures, vocal outbursts like laughter, etc.., that are aimed to be analysed by technologies of signal processing, or automatically generated by technologies of signal synthesis.

Computer systems and devices capable of sensing agreement, inattention, etc.., and capable of adapting and responding in real-time to these social signals in a polite, non-intrusive, or persuasive manner, are likely to be perceived as more natural, efficient, and trustworthy. In the context of assistance for living personal, social features seems to be crucial for acceptance of robot companion in a home environment.

2.2 Non Verbal Communication

2.2.1 From Cognitive Sciences to Human-machine interaction

Ekman in his book [14] uses three view points regarding non-verbal communication behaviours: the origin, usage and the coding. NV behaviours include emotions, attitudes, interpersonal roles and severity pathologies. The modalities used to convey each information vary. Hence, the face conveys more information about the nature of the emotions than the intensity of the emotional state. Body acts give information about the intensity and the nature of emotions. Postures provide information about the intensity of emotion and also about the overall affective state (pleasant or unpleasant feeling).

However there are also some epi-personal influences changing the expression of emotions and making it context dependent. The physical setting, the sex and the role of the person showing the act modify the meaning of the act, as well as the verbal context, voice and the tone. Even if the same gesture has a set of meanings dependent on the context, these meanings differ from the meanings associated to another act and can be recognized as a group of gestures sharing similarities. The decoding of a gesture can be difficult considering its occurrence and its variability both intra and inter-person. Even though most of the NV behaviours are subconscious and effortless, NV behaviours as well as verbal behaviour can be a lie, which also explains the difference among people and how they convey messages. This study of Ekman et al. on the inter cultural stability of NV behaviour encoding and decoding concludes that we can still extract invariant meanings from NV acts.

Ekman defines the usage of a NV act as the context surrounding the occurrence of the act. This context can be composed of several instances such as the external condition and environmental context, roles context etc.. Act and verbal behaviour relationship, which is the temporal correlation and meaning correlation between verbal and NV acts also takes part in the usage context. Other aspects of the usage are the internal feedback (awareness of emitting the act) and the intention of communication (deliberate use of NV act). In the context of interaction, external feedbacks such as reactions from the listener are an important aspect of the usage. Finally, the type of information convey by the emitter also takes part in the usage

aspect of the NV act.

An NV act can be of 3 origins according to Ekman [14]. It can be a reflex i.e. facial expression (neurological origin) stable intra species behaviour. NV acts can be inherited behaviours, acquired as species-constant experience (use your hand to bring food to your mouth). Alternatively it can be culturally or socially learned by imitation or reinforcement.

The third aspect of an NV act is its encoding and it can be done by three different code, following the classification of signs. First an arbitrary code, also called symbol, which express no correlation between the semantic and the act (hello and goodbye signs). The icons carry a clue, the acts looks like what it means (significant) in some way. Finally, the intrinsic code is used when the code IS its significant.

Argyle in his book «Bodily Communication» [15] repertories the different signals from different modalities used for non-verbal communication. The modalities considered by Argyle are facial expression, gaze, gestures & body movements, posture, contact, spatial behaviour, clothing, and vocalizations. He shows that recording of these modalities allows us to recognize the mood of a person.

P. E. Bull [16] follows the idea that communication implies a socially shared signal system or code. NVC is argued to be intentional or non-intentional, whereas verbal communication is always intentionally made to convey the verbal message. NV cues are valuable information in such a way that they allow to access information about the emitter that can be non-voluntarily transmitted (and hence show some more real intentions). Nevertheless, one can also control these non-verbal vectors of communication and hence the observer can be fooled. Bull claims the importance of posture and gesture in NVC where these channels have been neglected compared to facial features and speech cues. He gives some methods to study listener boredom vs. interest, and disagreement vs. agreement.

Jokinen in [17] describes properties needed by an interactive agent. He calls these properties Contact, Perception and Understanding. The feedback on the CPU enablements is important in smooth communication and is often expressed non-verbally. The agent must be sensitive to the relevant non-verbal signals.

Analysis of non-verbal communication signals is hence capital to the improvement of social signal processing and robot companion in human-robot interaction.

2.2.2 Intentionality in Human-Machine Interaction

The intention cues form a way of communication. Recognition of human' s intentions, goals and actions is important in the improvement of non-verbal human-robot cooperation. Intention recognition is define in [18] by the process of estimating the force driving humans actions based on noisy observations of human' s interaction with his environment. In this paper intention recognition is considered a discrete-time state estimation problem formalized in a Dynamic Bayesian Networks (DBN). Tahboub in [19] sees intention recognition as a substitution or complement to reliable and extensive communication which is a prerequisite for coordination and

cooperation. Indeed, in order to have a smooth interaction, intention recognition is essential. The DARPA/NSF in his final report on Human-Robot Interaction [20] recommends to improve the models of human-robot relationship and in particular to work on the intentionality issue. In [10] it is proposed to recognize intentional actions using relative movements of a human to a robot. Koo uses an IR sensor embedded on the robot to track and estimate the velocity of a person. He then infers intentional actions such as approach and departs using Hidden Markov Models (HMM) and position dependent model. This work seems not enough to estimate engagement. Indeed, one can slow down near the robot without wishing and interaction and in the other hand someone arriving fast near the robot might want to interact with it hastily. The relative position and speed are not the only features that can be used to estimate intentionality.

In his study Knight [21] points the importance for a robot to convey and hence to detect intentionality. It helps to clarify current activity and to anticipate the goals. Learning from the engagement from the human, the robot would be able to anticipate the interaction and also to learn adequate moment when, the robot itself can engage an interaction. In [22] engagement is defined as the process by which two (or more) participants establish, maintain and end their perceived connection during interactions they jointly undertake. Engagement is in the frame of connection that can be a collaborative task, spoken language, gestures etc. Sidney and Lee propose a model in three steps (1) initiation of interaction, (2) sustenance of interaction, (3) disengagement. As is presented in 3.5 we will see that the classification used in this project is based on the model.

2.3 Modalities for Non Verbal Communication and Intentions Recognition

The human body is equipped with an incredible number of sensors using different channels. These sensors are classified as proprio and externo sensors. In the proprioceptor ' s class, we find the temperature and equilibrium and in the externoceptor ' s class there are the visual, olfactory, and audio sensors. These externoceptors constitute the interface through which human perceive the world and hence the communication signals. If we consider the modality as the nature of the signal then in the context of human-human interactions, the modalities involved in the NVC can be seen as these 4 of the 5 human senses: hearing, sight, touch and smell. This modalities allow us to have more information than simply the speech. In a face-to-face interaction, most of the NV signals go through the hearing and the sight channels. From the hearing we can have gender cues, affective cues, agreement cues, mood cues and engagement cues. The sight gives for example clothing cues, age cues, affect cues (facial expressions, posture) etc.

If we consider now the modalities that can be involved in a human-machine interaction, most of them are through the manipulation of a device or through

sight and/or the hearing. The research in Brain Computer Interaction also allows treating physiologic signals. We detail the different modalities used in the social signal analysis in computer science research field. The modality channels through which non-verbal communication can be measured are the audio, face, posture & gesture, the physiologic aspects, clothing, gender, age etc.

2.3.1 Body Pose, Gestures and Proxemics features

A way of detecting engagement would be to consider only proxemics metrics. Nevertheless, it has been proved that in some cases these features can help in predicting the interaction. The classical features included in proxemics features are the relative position of the individual to the robot and its relative speed. For a collaboration to be successful, the distance between the robot and the human should be optimum and the speed controlled. As shown in Figure 2.2 there exists a social distance where some sensory sensations are not experienced during an interaction. It is important to know about such a distance so the person interacting with the robot does not feel uncomfortable. As an example, thermal sensor or haptic sensor used as a feed back for engagement detection would not be convenient for a companion robot.

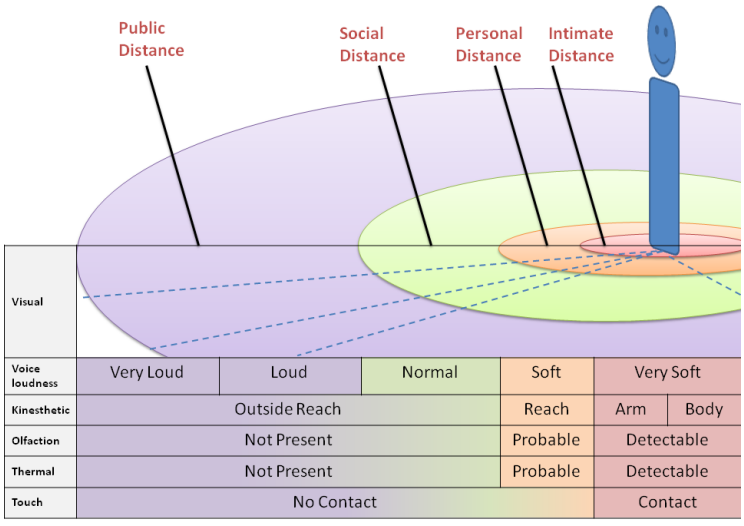


Figure 2.2: Hall's proxemic features [3]

Human spatial metrics can be useful measures to describe role, attention, and interaction. Psychologists have proposed many models to describe body pose metrics and their associated meaning. An overview of these metrics is presented in [3]. In [23], the authors propose a spacial model for a receptionist robot to infer the intentions of the human. The results given by this study coupling attention estimation and distance metrics are promising but the approach is limited to face-to-face

interaction. Glasnapp and Brdiczka in [9] propose a human-centered engagement model. They describe natural practices people use to develop and interact with an interactive display. Yet, the implementation and the validation of this model has not been proposed. Besides, this model is based on observations of the scene. To improve human-robot interaction, it is safe to consider that a robot-centred approach is more adapted, taking into account the mobility of the robot and the variability of the environment.
*0.4cm In [24], McColl proposes a 3D human body pose identification with model-based detection of the body parts. He uses a thermal camera, a time of flight range camera and a 2D camera in order to detect the skin information coupled with the pose estimation using the depth. This model-based approach is limited by the fact that only facial gestures are considered. Yet the classification of gestures is very interesting to integrate in a companion robot but we might consider using gestures as a way of giving direct orders to the robot. For that reason, we will consider natural posture and natural gestures only. As we have seen previously, there is no consensus on the meaning and the emotional characteristics of a posture. Psychologists such as Hall, Mehrabian [25] Schegloff [26] have proposed some metrics that have been used in computer assisted analysis of posture but there is no consensus on one particular model.

Posture is difficult to measure and evaluate using computer vision. Nevertheless, with the apparition of the Kinect Sensor from Microsoft Research [1] and other real-time 3D pose reconstruction, we are able now to evaluate the pose of a person. In a recent paper [27], the authors propose a coding system BAP to encode body actions and posture information in order to have a consensus. Unfortunately this coding system is applicable to a very sterile environment, using a frontal and a profile camera that doesn't suit human-robot affect detection.

2.3.2 Audio Features

Pantic in [28] and [2] lists some features into the audio signal that can be used to spot basic emotion such as happiness, anger, fear and sadness. It can be agreed on, that some audio features such as pitch, intensity, speech rate, pitch contours, voice quality and silence are good parameters to classify the emotional state of an individual.

In his paper, De Silva considers [29] in particular pitch and the pitch contours, in a bimodal emotion recognition context using audio and facial features. Chen in [30] also uses audio-visual fusion for multi-modal communication. In this technique, he uses pitch, intensity, pitch contours from audio features coupled video of the mouth motion in order to improve speech recognition and shows that an integration of both audio and video signal with an Hidden Markov Models classifier improves notably the speech recognition in comparison with the unmoral recognition. Considering the recognition of the engagement in an interaction, only few papers in the literature that uses audio features in a multi-modal frame. [31] pro-

poses an engagement estimator using head pose associated to audio features in a face-to face conversational agent interaction.

Even if we do not realize it we are often able to localize roughly a sound's source. We use the lag of the inputs between the two ears to estimate where the source is. Sound spatialization is not often used for affect detection. Some literature also invokes its interest in attention or focus estimation [32]. Indeed, often in literature, interactive agents using audio processing are involved in sterile face-to-face speech interaction. In this case, sound localization is unnecessary. Yet, if we consider an everyday life situation where a friend sees us in the street, running after us to attract our attention and engage in a conversation with us, sound spatialization would be of great use.

2.3.3 Facial Features

In terms of affect & emotion detection and speech recognition, a lot of studies have published results with a combination of face and audio features. De Silva, in his bimodal system [29] used optical flow to detect displacement and velocity of facial features tracking the mouth corners, the top & bottom points of the mouth and the inner corner of the eyebrows. Chen [30] detected lowering and raising eyebrows, opening eyes, stretching mouth, frown, furrow and wrinkles. Karpouzis used in [33] facial expressions and hand gestures to determine the emotional state. In his multi-modal system he used skin detection to locate the face, then he segmented the face before detecting its primary features while tracking the variation in distance between these feature points.

Concerning the engagement, the orientation of the head and the gaze seem to be crucial. Indeed as shown in [34] a speaker can be detected more easily with the combination of different features relative to the orientation of the face such as a mouth sensor. Face detection is already a first cue of interaction. The orientation of the face toward the interface is a sure sign of attention. Gaze tracking can give a better estimation of location of where the participant poses his attention, even though it is not perfect and can be ambiguous.

2.4 Fusion and Classification of Multi-modal Data

Each modality has a set of specific features and decision techniques that are commonly used to take state decisions in recognition. The difficulty in multi-modal data fusion is to select a method that would suit all the modalities monitored without the loss of information and with a gain of precision. There exist three levels of abstraction for multi-sensory data fusion considering at which point, from pure data to decision, we merge the information. Figure **??** presents the different levels on classical bimodal data fusion including audio and video recording. One of the issues of multi-modal recognition is to know how the sensory fusion will impact the

classification accuracy. Several hypothesis and techniques are nowadays used in affect recognition with multiple sensors. The first hypothesis is that the classification performance from multiple channels is super-additive, meaning that the classification performance from multiple modalities is superior to an additive combination of classifications from individual channels. In other words, *the whole is greater than the sum of the parts*. Another hypothesis is that there is redundancy between the modalities. This redundancy causes the fact that the addition of one channel to another channel constitutes a weak gain since the features of the two modalities are manifestations of the same spatio-temporal event. This section will first present

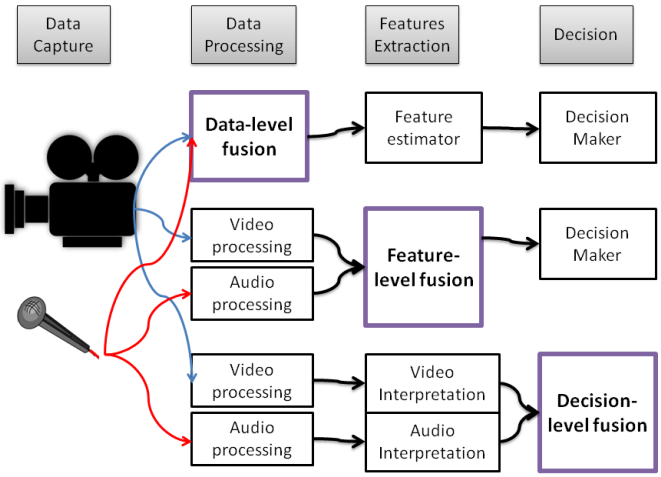


Figure 2.3: Abstract Fusion Level for Bimodal Sensing System [4]

each of the abstraction levels of fusion for multi-modal data, followed by a quick overview of the different machine learning techniques used in multi-modalities and compare the advantages and drawbacks of each method.

2.4.1 Fusion Level : Data

Data-level fusion is interesting when the correlation between the channels is strong. Indeed, data fusion often implies that sensors have been calibrated between each other. The integration at this level is made with raw sensory observations. As an example, one can image to fuse a depth image with a color image, and then to compute features on this RGB+D images. In this particular case the fusion implies that the homography between the depth and the color image. A drawback of this kind of approach is a possible loss of information due to early fusion (i.e. approximation of the homography matrix). Data-level fusion is mostly use for the fusion of same type of data but acquired from different viewpoints.

2.4.2 Fusion Level : Features

Fusion at the feature level aims to integrate the features extracted from the various sensors together before attempting to classify. Feature level fusion consists into merging the computed features from each modality into one cumulative feature vector, selecting the relevant features before feeding them into a classifier.

Perceptive systems have to deal with imperfect data and generate a conclusion such that the certainty associated with it varies accordingly to the input data. Feature-level fusion is less influenced by noise and sensors failures in comparison with the data-level. This sort of fusion level, assumes that the modalities are strongly coupled and synchronized.

According to [35], this type of fusion is considered more appropriate for closely temporally synchronized input modalities, such as speech and lip movements. This class of techniques utilizes a single classifier. To give an example of audio-visual integration using an early fusion approach, one simply concatenates the audio and visual feature vectors to obtain a single combined audio-visual vector, which finally feeds the recognition engine. The classifier utilized by most early integration systems is a conventional Hidden Markov Model (HMM) trained with the mixed audio-visual feature vector. Yet, in a first step and in order to build a less complex model, the use of Artificial Neural Networks (ANN) and Support Vector Machines (SVM) techniques is often made. In this vein, [36] uses a SVM classifier for the multi-modal prediction of degree of involvement in a conversation. Here the aim of the authors was to determine the valuable features to characterize, within a group, the involvement in a conversation of an individual. The use of SVM is a first step before the modelling of a temporal model for the prediction of the involvement.

If the available observations have to be sufficient compared to the amount of features, risking that the classification results become non-meaningful if not. On another hand, growing feature vector may stress computational resources. However, appliance of feature selection techniques may relieve both problems.

2.4.3 Fusion Level : Decision

The decision level fusion, also called late fusion, corresponds to the fusion of decisions computed for each modality (a subgroup of all the computed features). At this level, the states are first classified for each sensor and then integrated to obtain a global classification over all the modalities. Decision-level fusion allows integrating asynchronous but temporally correlated modalities. In a first time, a classification is performed on each modality independently, and then the decision is made by the fusion of the output of the mono-modal classifiers. This fusion can be executed by applying several criteria. Lingenfelser in [37] presents some of them.

2.4.4 Comparison of Fusion-Levels

The advantages of using late decision level are the computational cost of the training, from a bi-modality system going from $O(N^2)$ to $O(2N)$. The strict synchrony of the inputs is not required since they bring complementary information.

Despite important advances, further research is still required to investigate fusion models able to efficiently use the complementary cues provided by multiple modalities.

Synchronizing the video and audio channels and aligning emotional segments is a challenging task, especially with the complexity of mental states investigated which need temporal and spatial information to be captured [28]. Decision fusion gives a robust architecture and resistance to sensor failure. The approach however loses information of mutual correlation between the modalities. Indeed, there is no more temporal correlation between the modalities, and this is the main drawback of decision fusion techniques.

The fusion level depends of the purpose of the application of the multi-modal system. Indeed, if the multi-modality is used to compensate for a default of one of the sensor or the feature detected, then the decision level seems the best choice regarding the robustness and the fact that one can introduce a weighted confidence on the detection of each features. In another hand, if the multi-modality aims for capturing and analysing temporally and spatially related features and to build the decision considering this correlation then a lower level of fusion is necessary.

The variations in the frequencies of acquisition from different sensors used in multi-modal social signal processing make the fusion and synchronization hard at low and intermediate abstract fusion-level. Yet, fusion at decision level seems wrong if the aim is to accomplish human like multi-modal integration. Indeed, to do so, the input from the sensors cannot be considered as mutually independent but should be treated in a joint feature space depending on the context. In practice the context is very complex to sense and recognize. In addition, if the joint feature space is of high dimensionality, the format of the features varies, and the time is also considered, the problem faced is very hard to solve. In the literature, there are some examples of fusion of temporal multi-sensory data. Most of these approaches use learned probabilistic models such as Dynamic Bayesian Networks (DBNs) [38] [39]. Probabilistic Graphical Models such as DBNs and HMM have shown some good results in recognizing emotions, because they can handle noise, incomplete and temporal information thanks to probabilistic inference. However, it is also reported in [4] that DBNs approaches may fail when used to enhance complex behaviour. Indeed, such transitional models, handle easily small set of features and closely defined states.

In [36], the authors experimented several techniques of multi-modal fusion using Support Vector Machine Classifier to elicit the detection of involvement in a group conversational context. This study showed that the results using features fusion SVM classification had an improved accuracy.

3 A Corpus for Engagement with a Robot

As mentioned previously, a big part of the work accomplish during this project was to build a multi-modal dataset including interaction with a robot equipped with a Kinect device. Many aspects had to be taken into account such as device acquisition optimization, temporal synchronization, disk space and storage of the data. This chapter presents the method used to build the dataset needed to test our hypothesis. The first section 3.1 explicits the need to build a dataset that we encountered in order to test our hypothesis. The section 3.2 introduces the hardware used to record the corpus. The following section 3.3 discusses the criteria needed to build a social signal scenarios that are realistic. Section 3.4 deals with the experimental choices and constraints of the recording. The steps of the interaction model are depicted in section 3.5. Finally the section 3.6 presents the scenarios included in the dataset and the section 3.7 shows some samples of the recorded data.

3.1 Need for a Dataset

In the frame of a companion robot, we want to work with consumer devices and in a natural environment. Even thought the tendency is to spread more and more physiological sensors such as R. Picard's pulse bracelet Cardiocam, nowadays, physiological signals are still invasive and too expensive for the users to be release wildly. The physiological modality is not considered in this work, yet it might be enriching to include it in further work. The modalities recorded the audio, facial, posture and spatial positions modalities, are relevant taking into account that we want the system that is individuals invariant (working for any user).

In order to evaluate this work, the hypothesis has to be confronted to data. In the context of robot companion, the sensors considered are the one commonly on such robots, microphones, video sensors, depth sensors, lasers etc.. There exist datasets in the field of social signals processing dealing with non-verbal communication using multi sensors. The datasets available for affect recognition are unfortunately more often for face-to-face interaction with persons sitting and interaction with the speech only. The SSPNet association proposed the SEMAINE-DB dataset [40] where several person have been recorded in a face-to-face speech interaction. This database is suitable for a desktop environment with interaction with virtual communicant agent. Unfortunately, this dataset suits less human-robot interaction and especially if the non verbal cues of social signal that are involved in the engagement of interaction are more diverse than the facial expression and the speech characteristic. Other datasets exist that uses the Kinect sensors and 3D informations such

as in [41] which presents a Cam3D dataset centered on facial and hand movement associated with audio recording. Yet, the proposition of a robot centred dataset for multi-modal social signal processing has not been proposed. The limitations of the existing multi-modal datasets regarding the current subject of these researches are set and the proposition of a new dataset is made. For these reasons, the sensors equipping the Kompai robot have been used to build a dataset where the interaction is a physical interaction with the robot. The dataset proposed would be in the robot view-point and composed on various sensors. The scenarios included in this dataset are presented in this chapter. Enrichment of the dataset are proposed for further work, including different modalities of interaction with the robot companion, in order to improve the efficiency and the robustness of the detection.

3.2 Hardware Sensors

3.2.1 Kinect Sensor

This project started before the release of the new Kinect Sensor for Windows [1]. Hence we used the Xbox360 Kinect sensor. The Kinect sensor is composed of several components some of them are presented in the Figure 3.1. The advantages

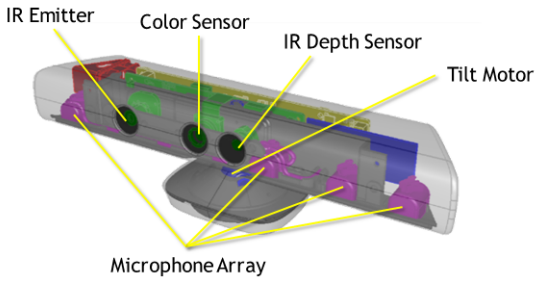


Figure 3.1: Components of the Kinect Sensor [1]

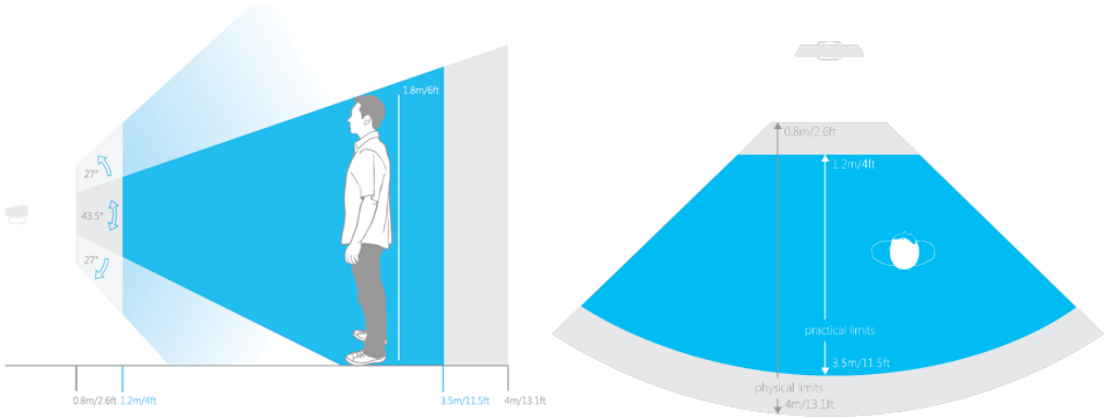
of using such a sensor is the consumer price of this device and its growing utilization in computer vision assisted system. Nevertheless, some difficulties are raised when full use of the device is required for real-time acquisition of all the sensors equipping the Kinect. Indeed, if a query of a modality of the Kinect is made individually, the frame rate reached for acquisition is close to the maximal one (30FPS for the video, the skeleton and the depth channels). Beside this, a way of representing and storing the data had to be found. Hours of recording were needed to build the dataset, yet the wish to keep the data as pure as possible was made and some optimization to store big quantities of raw files had to be implemented.

Depth Camera (using Infra-red laser) The depth range is limited from 80 centimetres to 4 meters (Fig. 3.2b). The accuracy of the measurement in this range is within 2 millimetres. The depth is measured in meters from the camera along the

Z axis as shown in Figure 3.2. The X and Y axis are measured in pixel coordinates.

The resolution of the depth image is of 320x240 pixels by default and can be set at maximum at 640x480. The depth camera horizontal field of view is of 58.5 degrees.

The frame rates varies according to the services queried on the Kinect Sensor and is at maximum 30 frame per second. One of the difficulties encountered during the data recording was that all the sensors of the Kinect were queried, hence the frame rates of acquisition had a tendency to decrease. The challenge was to keep a frame rate close to the maximum by optimization of the code for acquisition and using high performance hardware.



(a) Vertical field of view (b) Horizontal field of view

Figure 3.2: Vertical (a) and Horizontal (b) field of view [1]

Skeleton Tracking The Kinect for Windows supports up to two skeletons being tracked at the same time. Yet, up to six players can be detected. Only the skeletons tracked (with more confidence) are stored in the dataset. A Json structure is proposed to serialize and map the skeleton's frame informations.

RGB Camera (for visible light images) The resolution of the RGB image is of 640x480 pixels by default and can be set at maximum at . The RGB horizontal field of view is of 62.0 degrees.

Tilt Mechanism and an Accelerometer The camera tilt varies of ±27 degrees and can be requested at maximum 15 times in 2 minutes.

Microphone Array The array is composed of four microphones aligned. As shown on the Figure 3.3. It is calibrated to give an angle of the position of source of a detected sound with a certain confidence in the Kinect reference frame. It also outputs the beam the more stimulated by the sound source.

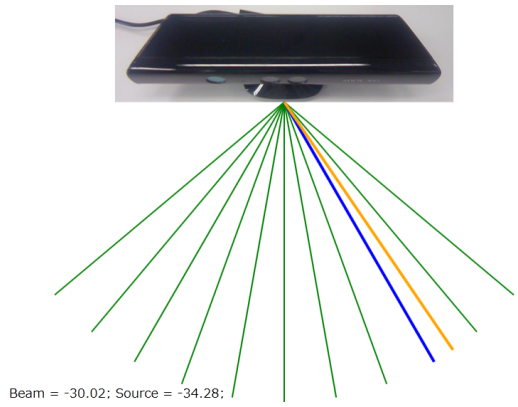


Figure 3.3: Kinect Sound Source Localization [1]

3.2.2 Kompai robot

The Kompai robot as been loan to our lab for few weeks by our partner Robosoft [5]. It allowed us to record a corpus of data. The Kompai robot, Figure 3.4, released in 2010 aims for help to elders and dependent persons. It is composed of a RobuLAB mobile platform containing the wheel actuators, obstacle detection system, manual remote control utilities etc. The mobile platform is topped by a tablet serving as interface with the user, a Kinect sensor added by ourselves on the robot, a pair of microphones, a motorized camera and a speaker device. As mentioned before this work, the Kompai robot is the robot hosting the sensors that are used for the engagement sensing. As described in [5], Kompai is equipped with many sensors, in this work only the nine Ultrasound sensors and the sixteen Infrared sensors are monitored. These sensors are positioned on the base of the robot as shown on Figure 3.4. The range on detection of the Infrared sensor is from 20cm to 150cm. A camera placed on the top of the robot is also used to record videos during the experiment.

3.3 Realistic Dataset

As explained previously in this chapter, this project needed the elaboration of a new multi-modal dataset in order to test our approach. This section deals with the thought process that as been followed in other to build a realistic dataset.

The common dimension of all the modalities is the time (chronology of events). The key of the data fusion is to correlate the different modalities on the same time scale. This synchronization of events, developed in section 4.2, is made possible by tagging all the capture data with relative or global time.

R. Picard in [6] states five variables that may affect data collection. The first factor is the spontaneity of the expressed emotion. The emotion can be either elicited by a stimulus or asked to elicited (activated or acted). Another influence can come from the environment of the recording, and the question here is that are

Figure 3.4: The Kompai Robot, Robosoft [5]

the emotions expressed and recorded similarly in a lab setting and in a real-life situation? Next question to be considered when recording affective data is: should the focus be on the expression of the emotions or on the internal feeling? The internal feeling would be measured by retrospective interviews of the participants. The awareness factor of the recording is another factor. Indeed, what is the influence of open-recording in comparison with hidden recording on the recorded data? Finally, should the emotion be presented to the subject as the purpose of the experiment or not?

Regarding this project's matter, the engagement is relatively spontaneous. It is asked to the participant to interact, yet its intention toward the interaction cannot be elicited artificially. The intention will show whenever the participant plan to interact. The participant is explained that what is measured is its reaction while playing the game. The goals of measuring intentions is still hidden, there is no awareness to the recorded factor by the participant. The recording is made in a smart environment, similar to a flat. Yet, for many of the participant this room is new and this can create some fluctuations in the behaviours. Since this work focuses on the external expression of the engagement, no questionnaire evaluation were performed for this experiment.

The data recorded to build the corpus are presented in the table 3.1. Only some of them where analysed for to extract features for the engagement detector.

Data	Sensor	Maximal Frame Rate
Telemeters distances	Kompai's Laser	12.5Hz
Ultrasound distances	Kompai's Ultrasound	12.5Hz
Audio	Kinect	16kHz per channel
Sound Source Beam and Position	Kinect	8Hz
Skeletons	Kinect	30Hz max
RGB Video	Kinect	30Hz max
Depth Video	Kinect	30Hz max
RGB Video 2	Webcam	15Hz
Button Press	Tablet	-

Table 3.1: Shows the different data collected in the dataset, their corresponding sensors and the optimal frame rate of acquisition

The next section will present which of them where used and what features where obtained.

In order to test one of our hypothesis which is that the position of the person is not enough to detect an intention of interaction, we propose to make in practice scenarios where the user passes close to the robot but with no intention of interaction. The robot is for this work immobile, but since all the features computed are robot centred, we can imagine that in further work our system would be able to detect the intention and anticipate by turning toward the user or signalling its availability to interact.

3.4 Experimental implementation

The interaction in this dataset consist in a small flash game. The participant uses a stylus to click on ground hog coming out from holes. The parameter tested is actually with pre-interaction cues and without pre-interaction cues but in a real-life manner. The hypothesis is that the interaction phase is preceded by a pre-interaction phase where the participant shows some signals of its intention of interaction. The hypothesis goes further with the assumption that these cues are detectable with the sensor that equipped our version of the Kompai robot.

Of course in real life any individual don't express these signals the same way. Some variability has to be introduced in the pool of participants. The participants are from 20 to 35 years old and are female and male. The voice, clothing, posture varies among the participants. The testing data are taken from different sessions of recording. The duration of the interaction also varies from 2 to 10 minutes according to the participant will.

The interaction time interval is made easy to detect by labelling the beginning and the end of user clicks bursts. Hence the pre-interaction phase preceding the beginning of interaction (first click) by registering the time in which the participant uses the stylus and touches the screen. Labelling the data is usually a fastidious

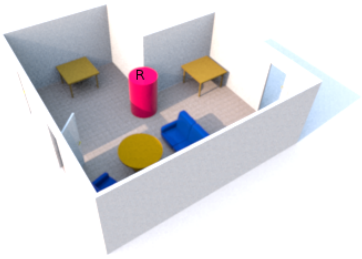
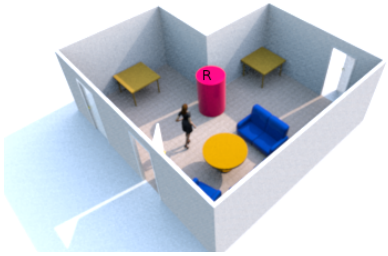
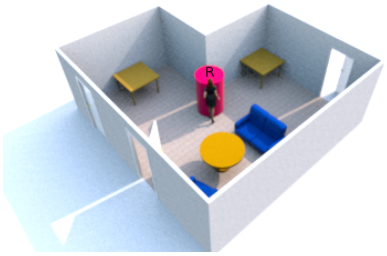
NONE	
NO-ONE	SOMEONE_AROUND
	
WILL_INTERACT	INTERACT
	
LEAVE_INTERACT	
	

Table 3.2: The different stage considering intention of interaction. The red cylinder symbolizes the robot

task but very important to be able to train the system correctly. The use of the click data on the tablet allowed us to make this task automatic and faster than manual labelling.

3.5 Steps of the interaction process

The process of interaction as been describe by Sidner and Lee in [22] proposed a model in three steps : (1) initiation of interaction, (2) maintain of interaction and (3) disengaging.
This work follows this approach by modelling the events as :

WILL_INTERACT (1)

INTERACT (2)
LEAVE_INTERACT (3)
NONE (the rest of the events not related to interaction with the robot)
These events are illustrated in Table 3.2. Located in the homely-like environment the robot is placed at the cylinder space marked with a R on the figures. Another segmentation of the interaction scenarios with one more event is proposed by spiting the NONE event in two as followed:
WILL_INTERACT (1)
INTERACT (2)
LEAVE_INTERACT (3)
SOMEONE_AROUND
NO-ONE
The SOMEONE_AROUND event is labelled when someone is detected in the room but with no wish of interacting with the robot. When the nobody is in the room, it corresponds to the NO-ONE event.

3.6 Scenarios

The data are recorded within two different scenarios performed several times by different participants in a smart environment where the Kompai robot is placed. This monitoring is made by the sensors that equip the Kompai robot. Each participant is given randomly one or several actions to perform in the room. The room is similar to a small flat (Figure 3.5). It is asked to the participant to enter the room by different doors, perform some realistic actions and going out. One of the actions is to interact with the robot. The other actions were walking, sitting, or poring water from the sink.

The presence of interaction will be measured by the time delay in which the tablet is stimulated by the stylus click. This measure gives us a time interval for the interaction, here the intensity of the interaction or on the mood of the participant during the interaction is neglected, since the focus is made on the intention of engagement. According to this variation on the presence of the interaction, the presence and the intensity of the pre-interaction cues have to be sensed. To randomize the attributions of the actions for the participants is also a way of controlling certain pattern in the parasite variables that can appear when experimenting with real data.

3.6.1 Scenario 1 : Passing By

In this first scenario, one participant is ask to go through the room twice by different doors (A), (B) or (C). The Figure 3.5 shows the setting of this scenario.

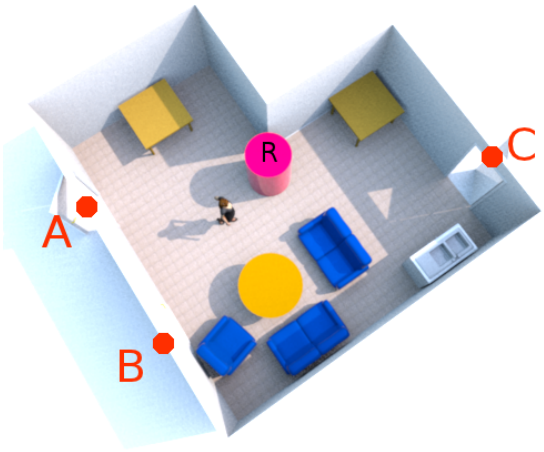


Figure 3.5: Scenario 1, Passing by

3.6.2 Scenario 2 : Playing cards together

In this second scenario, 3 persons are asked to start a card game in the living-room part of the flat. A telephone placed in the room is used to ask one of the participant

Figure 3.6: Scenario 2, Playing cards together

to execute an action (interaction, or using the sink for example). The Figure 3.6 shows this scenario when one of the participant is entering in the room while the other two are already sitting.

3.7 Samples of the corpus

The Table 3.3 shows some sample of the dataset that we propose for 4 of the events WILL_INTERACT, INTERACT, LEAVE_INTERACT and SOMEONE_AROUND. The first row pictures the moving object detected with the telemeters data. The second row is extracted from the Kinect video camera. As one can see the event LEAVE_INTERACT is quite confusing with the INTERACT state, this is explainable by the fact that this event is very short and so the number of multi-modal frame corresponding to this event is only of few .

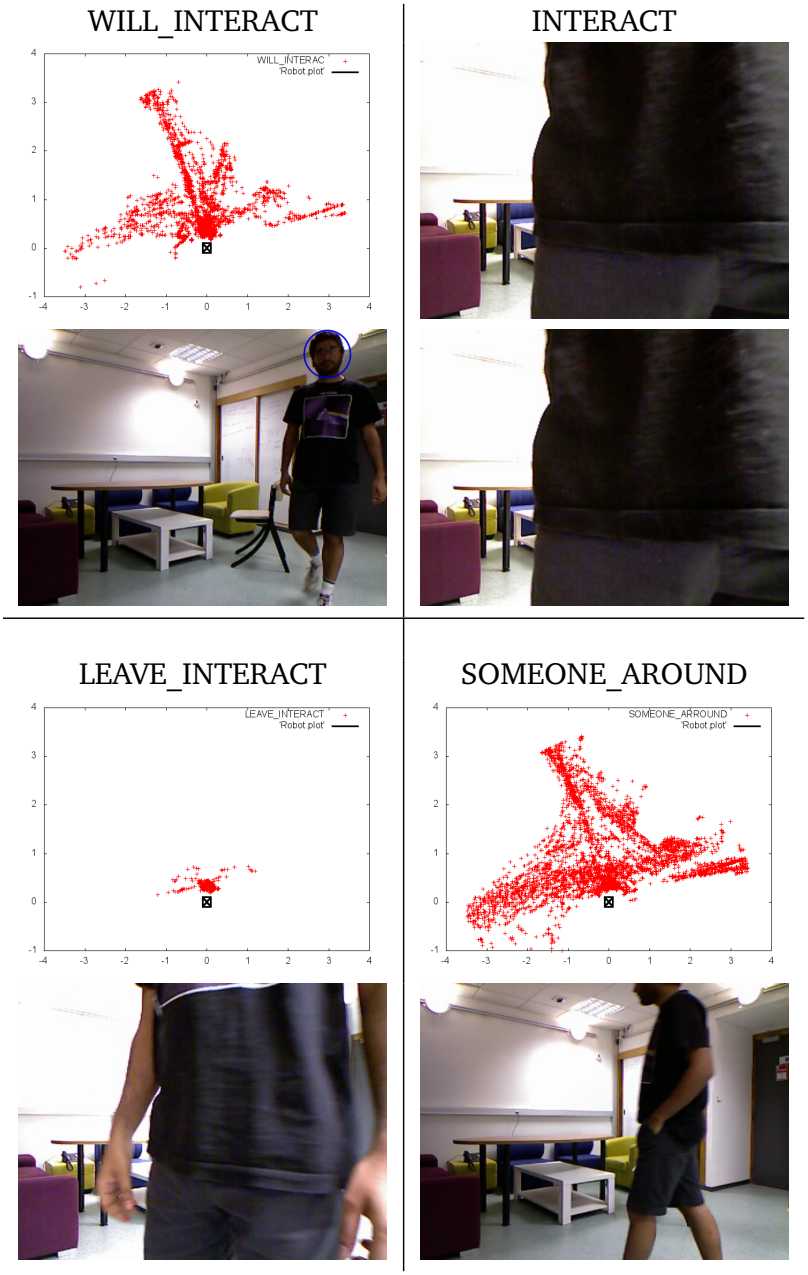


Table 3.3: Samples of data recorded with the Kinect sensor. First row without intention of interaction, Second row with intention

The dataset in numbers The ratio of frame for each event is not equivalent. The Figure 3.7 shows the proportion of each event in term of percentage of total number of frame recorded.

The recording of the corpus has been made during three sessions of one to two hours of acquisition of data. In total, the corpus includes 29 acts of interaction with the robots, made by 15 different participants. A variability of the gender of the

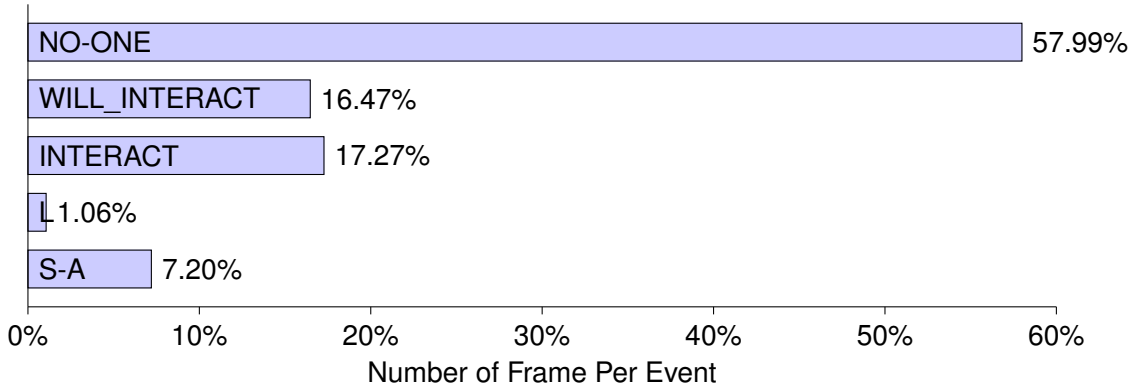


Figure 3.7: Repartition of the events on the frame numbers, with L : LEAVE_INTERACTION and S-A : SOMEONE_AROUND

participants have been preserved. The total size of the uncompressed data set is around 300Go.

4 Engagement Features Extraction and Synchronization

In order to characterize the engagement, some features were extracted from the corpus previously introduced. In order to have less computational time of treatment we choose to first compute the features needed for our matter and then to synchronize them with a unique time scale. The section 4.1 presents of the different techniques of features extraction used in the frame of social signal processing. The synchronization methods and the labeling of the computed features is then explained in part 4.2.

4.1 Features extraction

The feature algorithms are all noisy, and one of the interest of the multi modality is to be able to compensate for some default of some features. The output of the treatment of the data is for each frame a vector of 32 attributes coming from the treatment of some modalities of the dataset. The modalities used in the frame of the intention of interaction detection are the laser **??** that allows position and speed estimation, the audio 4.1.2 used for speech detection and sound localization, the Kinect skeleton information 4.1.3, and the Kinect video 4.1.4.

4.1.1 Laser

Position of moving object As mentioned previously, using proxemic features is of classical use to determine the intention of interaction. Indeed, as presented in [10] the relative position of the user in the robot reference frame informs on his engagement. Once again this features is not sufficient to detect the intention of interaction for robot in homely space. The choice is to integrate this features in a multi-modal detector to compensate for misleading interpretation. The laser sensors that equip the Kompai robot gives every 80ms, the distance over 270 beams, with a range from -135 to 135 degrees. A background subtraction using the mean of the telemeters value is used to detect moving objects in the room. The distance and the beam associated to the detected moving object are used to compute the position x and y into the robot reference frame. The trajectories of the moving object are showed in the Table 3.3. The stream of the laser begin the steadiest and one of the longest, it is used as frequency for the whole multi-modal system.

Estimation of Speed A Kalman Filter is used on the set of moving objects detected by the laser to compute the speed and the acceleration of the moving object at each frame. The Kalman filter is an iterative prediction estimation algorithm allowing to introduce measured data (in this case the position x and y of the moving object) and to estimate dynamics such as the position, and speed in the two dimensions. The implementation of the Kalman Filter over the telemeters moving object data as been made using the OpenCV library in C.

4.1.2 Audio

The microphone array embedded in the Kinect sensors is a four-element linear microphone array, using 24-bit ADC and providing a local signal processing (acoustic echo cancellation, noise suppression).

Speech detection Taking the audio stream coming from the four microphones the Speech Audio Detector (SAD) provided by the PRIMA team in [42] outputs the time intervals of voice activity detection and the highest pitch within each interval. The SAD detector is used to label the audio stream. Hence for each multi-modal frame we can retrieve if the audio was voiced or not. The Figure 4.1 shows a section

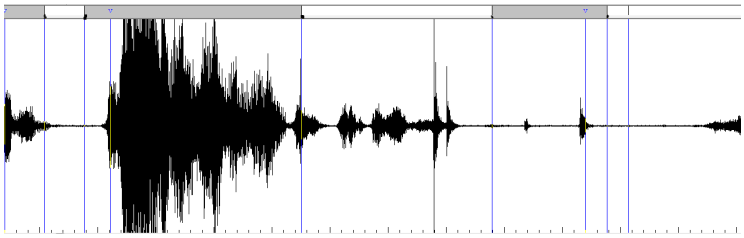


Figure 4.1: Speech Audio Detector Labelling the Audio Stream from the Kinect Sensor

of audio output from the Kinect sensor which presents some voiced events labeled on top by the gray bar.

Sound Source Localization The sound source localization is an interesting features, because it gives an angle from where a sound was emitted. The sound in this case can be a voice or the steps of the user coming in the room. Unfortunately the Kinect Sensors SDK does not provide information about the calibration of the microphones such as the relative position of the microphones. For these reasons, the technique used for sound localization is the one provided by Microsoft Kinect SDK that uses the beamforming technique. This beamforming supports 11 fixed beams placed within a range from -50 to +50 degrees in 10 degrees increments. The Figure 3.3 shows the repartition of the beams. The source localizer outputs the stimulated beam (rough estimation) and the source position (more accurate angle) associated with a confidence. The frame rate of the localizer is of 8Hz.

4.1.3 Kinect Skeleton

The Skeleton tracking is one of the attractive features proposed by the Kinect sensor. Indeed, it allows to to real time pose and gesture recognition. A part of the theory used to recognize the skeletons is presented in [43] by the Microsoft Research Cambridge and Xbox Incubation team. The methods results in using the depth map from the Kinect Infra-red sensor to segment silhouette from the background. A machine learning process using randomized decision forests is then used to train the system and to be able to propose a set of joints. The skeletons provide information about the 3D positions of 20 joints of the body in the Kinect reference frame.

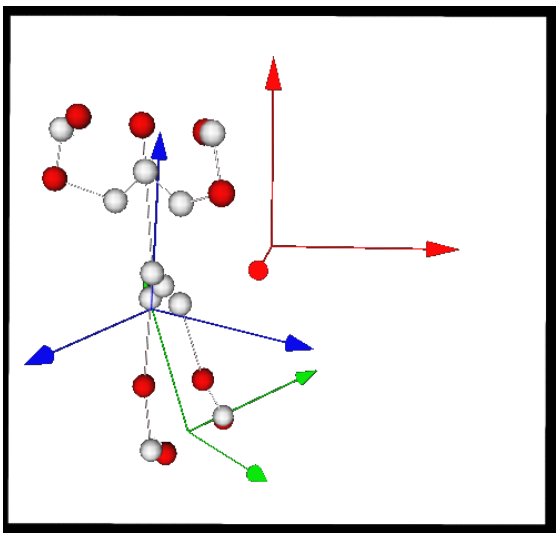


Figure 4.2: Stance and Hip pose and torque, body pose features computed from the skeleton informations of the Kinect sensor

Body pose As presented in the previous Chapters body pose features give indications on intention of interaction. Indeed, this features is also used to measure the level of engagement of a user into a task as in [44], proposing a measure of the Body Lean Angle. The body features that are proposed in this project are based on Schegloff metrics presented in [26] and in [3]. This features aims to depict the body pose of the individual. In this work, the accent is posed on the stance, the hips, the torso and the shoulders positions and orientation relatively to each other. The first body feature is the position and orientation of the feet represented by the stance pose and torque. The Figure 4.2 shows the stance and hip poses and torques computed in real time using the tracked skeleton input from the Kinect sensor. On this Figure the coordinate frame of the Kinect sensor is pictured in red. The hip pose and hip torque features represent the hip position relative to the stance pose and torque. In the same way, the torso pose and torque are relative to the hips and the shoulder features relative to the torso. What is interesting in computing these

body features is that they depict the orientation of the body part relatively to the Kinect sensor placed on the robot. We propose an implementation of the Schegloff's metrics using the Kinect sensors in order to characterize the body pose. Intuitively, if the interaction is long and focuses the attention of the user the body pose will reflect this engagement by being stable (two feet on the ground), facing the robot (hips, torso and shoulder oriented face to the robot) etc.

Distance A distance associated to the skeletons positions is also computed using the average z-value of several joints of the skeletons.

4.1.4 Video

The video extracted from the RGB camera of the Kinect sensor is fixed on the Kompai robot. From this video data we propose to detect frontal faces. We used

Figure 4.3: False Positive of the Haarcascade Face Detector

a trained machine learning system using Haarcascasdes method. The training is provided by the OpenCV library [45]. The Haarcascasde implementation is based on the Viola-Jones method with some optimization to make it real-time by using the cascade principle. The Figure 4.3 shows the quasi constant noise with a false face detection.

4.2 Synchronization and Labelling

In order to be able to fusion the data, we compute first the features on each of them and then merge the features in a matrix with number of columns equal to the number of features and number of line equal to the number of multi-modal frame recorded. The multi-modal frame rate is set on the frame rate of the telemeters features input. This choice has been made for two clear reasons: the need to compare the detector with the telemeters only with the multi-modal detectors and

the fact that the telemeters have a constant frequency of output, with one of the lowest frequency from the set of sensor that we capture from, allowing us to not have any artefact with interpolation of data within the time between two frames.

M. Anne in [46] describes a confidence in the input from a sensor. According to the author's model, the confidence on the new input decrease with the time delay between this new input and the previous one. In our case the time delay between two frames being relatively short, this confidence remains high.

The labelling of the dataset with the 5 classes (WILL_INTERACT, INTERACT, LEAVE_INTERACT, NO-ONE, SOMEONE_AROUND) is made using both the tablet touching informations and the telemeters. The INTERACT time-interval is labelled from first touch of the tablet till the last click. WILL_INTERACT events correspond at start to the entrance of a person in the room when an interaction is following till the beginning of the interaction. The LEAVE_INTERACT is the 5s following every interaction. The NONE event, including NO-ONE and SOMEONE_AROUND events, correspond to the rest of the time, respectively, there is no-one in the room or a person is present but there will be no interaction.

In order to synchronize the monitored data from the different modalities, it is needed to have a time of acquisition of each input. The data collected through the Kinect sensor such as the skeletons ṕositions, the video and the depth are tagged with a time relative to the Kinect sensorś initialization. The laser data are labelled with a absolute time stamp thanks to the real-time micro-controller. According to [4], in general, rapid behavioral signals can be recognized from 40-ms video frames and 10-ms audio frames. The frequency of the telemeters input is the most regular and stable on, hence it is used as multi-modal frame rate. This frame rate allows us to take the current value for each data and to not interpolate. This multi-modal interval between each frame is then set as 80-ms.

5 Experimental Multi-modal Fusion For Engagement

In the frame of this work, this first step is to test all the modalities that can help to detect intention of interaction. Then, a selection can be made among the most relevant multi-modal features. The evaluation of this detector deals with the gain in the detection of the intention of interaction by using multi-modality rather than simple position information.

The choice has been made to use a feature-level fusion as presented previously in 2.4.2. Indeed, the data being of different types the data-level fusion as been discarded. The aim of having a robust and general proof of the multi-modal impact on the recognition of engagement also led us to not consider the decision-level of fusion. Indeed as presented in 2.4.3, the decision level techniques are quite heuristic, less automatic and less natural than the feature-level fusion approaches.

Two tools were used for the classification, the Scipy library through Sklearn [47] and the Weka toolbox [48]. The technique used for the classification are the Multi-class Support Vector Machine (SVM) from Sklearn and the Artificial Neural Network (ANN) technique from Weka. A trial with Hidden Markov Models has been attempted but the results where not conclusive probably because the prior transitions learning was lacking. A model of the system with a DBNs system as been experimented, yet because of the high number of features and the limited number of instances, the results were noisy and not conclusive. The delay to record the dataset constrained the number of sessions of monitoring. The first results show that more training data would be valuable to include the temporal aspect of the detection, yet the classification using Neural Network (section 5.2.2) and Support Vector Machine (section 5.2.1) allows us to conclude in the favor of our hypothesis. The number of feature for each time-frame is of 32 for 155879 instances. There is no equi-repartition of the data for all the events.

The number of features J is going from 0 to $J = 32$. Considering the instance index I is going from 0 to $I = 155879$, we express a features of an instance as $f_{i,j}$. The matrix of data $M_{I,J}$ is shown in the following where each line is a time-frame of data considered as a feature vector :

$$M_{I,J} = \begin{array}{c} \begin{array}{ccccccc} SAD & beam & \cdots & speed_x & \cdots & shoulder_{torque} \end{array} \\ \begin{pmatrix} f_{1,1} & f_{1,2} & \cdots & f_{1,j} & \cdots & f_{1,J} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ f_{i,1} & f_{i,2} & \cdots & f_{i,j} & \cdots & f_{i,J} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ f_{I,1} & f_{I,2} & \cdots & f_{i,j} & \cdots & f_{I,J} \end{pmatrix} \end{array}$$

This chapter explains how we pre-processed the dataset in section 5.1. Section 5.2 detatils the classification techniques used and the results from this training/tetsing loops.

5.1 Prepare the dataset for the Classification

k-Cross Folding In order to train a model and to test it afterwards, the dataset needs to be split in two, a training set and a test set. A way to randomize this splitting is the k-cross folding. In this method, the dataset is partitioned in k subset within one is kept for testing and the $k-1$ others are used for training the model. This splitting process is repeated k times so that each subset is used once for testing. K-cross validation allows to be sure that the splitting is quite random. Since the events are not equiproportionate and temporally related, we used a *stratified* k-fold-cross validation that keep the same proportion of the different classes in the splitting process.

Dimensionality reduction A trial of dimensionality reduction of the feature space has been made using the Principal Component Analysis (PCA) and the Linear Discriminant Analysis (LDA) using the Sklearn tool-kit. The results were not conclusive, the reduction of the dimensionality gave strictly the same performance during the classification where we were expecting an improvement. We are still investigating this aspect.

The Minimum Redundancy Maximum Relevance technique have been performed in order to highlight the best features for our detection system. This dimensionality reduction technique has the advantage compare to LDA and PCA of giving the more relevant features instead of building new features by combination of the observed ones. Hence it could allow eventually to discard some less relevant features in order to optimize the detection of engagement process.

5.2 Classification Results

In order to classify our features, we chose to use two kind of classification. Many other techniques could have been applied, but in order to have a comparison between the classification and between the learning with multi-modality or with telemeters only, we decided to test the ANN and SVM techniques. For this two techniques we built and tested two classifiers one for the multi-modal dataset (including the whole 32 features) and one for telemeters dataset (a subset of the multi-modal dataset including the LIDAR moving object and tracking informations only).

5.2.1 Neural Network

The Artificial Neural Network is a graphical layered model. We used the Weka toolbox to perform this classification. The use of ANN is common to infer model from observation. In our case, we suppose that our feature can characterize the engagement, the use of ANN technique can help us to test this hypothesis. ANN is a good classifier to build prospective detection especially with large feature vector. The test results of the ANN classification are presented in the Table 5.2 for the telemeters, and the Table 5.1 for the multi-modal dataset.

Class	Precision	Recall	FPR	Accuracy
No-one	0,95	1,00	0,07	0,97
Will Interact	0,90	0,87	0,02	0,96
Interact	0,84	0,95	0,04	0,96
Leave Interact	0,21	0,01	0,00	0,99
Someone around	0,76	0,41	0,01	0,95
	0,91	0,91	0,02	0,96

Table 5.1: Results of Multi-Modal Neural-Network 5-class classification using Weka

Class	Precision	Recall	FP-Rate	Accuracy
No one	0,95	1,00	0,08	0,97
Will Interact	0,91	0,77	0,02	0,95
Interact	0,77	0,96	0,06	0,94
Leave Interact	0,00	0,00	0,00	0,99
Someone around	0,75	0,35	0,01	0,94
	0,90	0,90	0,03	0,96

Table 5.2: Results of Telemeter Neural-Network 5-class classification using Weka

First, these results show that the overall precision and recall of the classifier for our classes is slightly better in the multi-modal approach. Concerning the engagement class, WILL_INTERACT, the system returns more relevant event as an engagement in the case of the multi-modality and its accuracy is improved. For the engagement detection, in a practical point of view, the accent has to be put on the good performance in term of recall and a low false-positive rate. The Neural Network classifier gave better recall rate in multi-modal.

5.2.2 Multi-Class Support Vector Machine

The results of the 5-classes classification using support vector machine for the multi-modal data are presented on Table 5.3. For the telemeters classification the results are presented by the Table 5.4. We see comparing this tables that the precision

and recall scores for WILL_INTERACT class are significantly improved by the multi-modality. Also for this same class the False-Positive rate higher in the case of the telemeters only. The aim of this detection was especially to decrease this rate of misclassifying an event as WILL_INTERACT, hence the system has less chances to predict an interaction when there will not be one and to disturb a user with no intention of interaction.

Class	Precision	Recall	FP-Rate	Accuracy
No one	0,92	0,88	0,11	0,89
Will interact	0,92	0,71	0,01	0,93
Interact	0,54	0,77	0,15	0,84
Leave interact	0,04	0,10	0,03	0,96
Someone around	0,52	0,29	0,02	0,93
	0,78	0,78	0,06	0,91

Table 5.3: Results of Multi-Modal SVM 5-class classification using Sklearn

Class	Precision	Recall	FP-Rate	Accuracy
No-one	0,68	1,00	0,65	0,72
Will interact	0,80	0,68	0,05	0,90
Interact	0,00	0,00	0,01	0,81
Leave interact	0,00	0,00	0,00	0,99
Someone around	0,76	0,01	0,00	0,93
	0,69	0,69	0,09	0,87

Table 5.4: Results of Telemeter SVM 5-class classification using Sklearn

5.2.3 Minimum Redundancy Maximum Relevance for Features Relevance

In order to evaluate the relevant features for the multi-modal detection of engagement we used a MRMR dimensionality reduction of our feature vector of various amplitude before performing the SVM learning. The MRMR algorithm is proposed by [49]. The Figure 5.1 shows the impact on the precision of the steady feature reduction. The precision drops when only six features are taken into account, yet it remains pretty stable and even slightly increases along the feature reduction. These results confirms the fact that there are many correlations in the feature space. Yet some of these features seems to be fundamental for a better detection and to keep a precision higher than the telemeters' one.

Equivalent conclusions can be made on the Figure 5.2 regarding the recall performances.

The relevant features for the engagement detection highlighted by the MRMR process are presented on the Figure 5.3. This Figure presents, for each features their

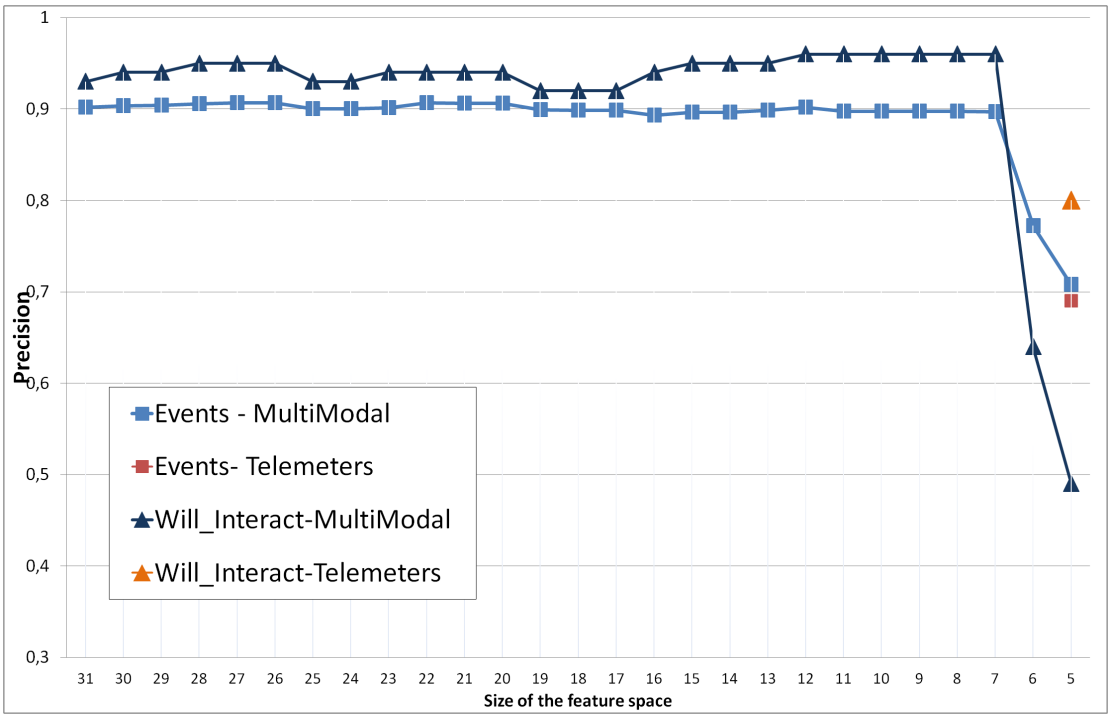


Figure 5.1: Precision evolution with the decreasing number of multi-modal features in comparison with the telemeters for all the events and for the event WILL_INTERACT

selection rate while constricting the feature space to the most relevant features and the one carrying the less correlations. The first remarks on these results is that the 7 highest rated features are coming from heterogeneous modalities. The *face_size* and *face_x* are respectively the relative size and position of the face in the video of the Kinect. The *beam* and the *angle* are the sound localization features from the Kinect's microphone array. The *shoulderPose_rot* correspond to the relative orientation of the shoulder in the body, and is extracted from the skeletons informations. The telemeters information are considered as relevant, with the high selection rate of the speed *target_vx* and position *target_y*.

From these results, some intuitive aspects of the engagement recognition are comforted. Indeed, the importance of the body pose, such as the orientation of the shoulder is exposed. Also the position of the face in the centre of the image shows that the person is facing the robot which a priori shows its engagement.

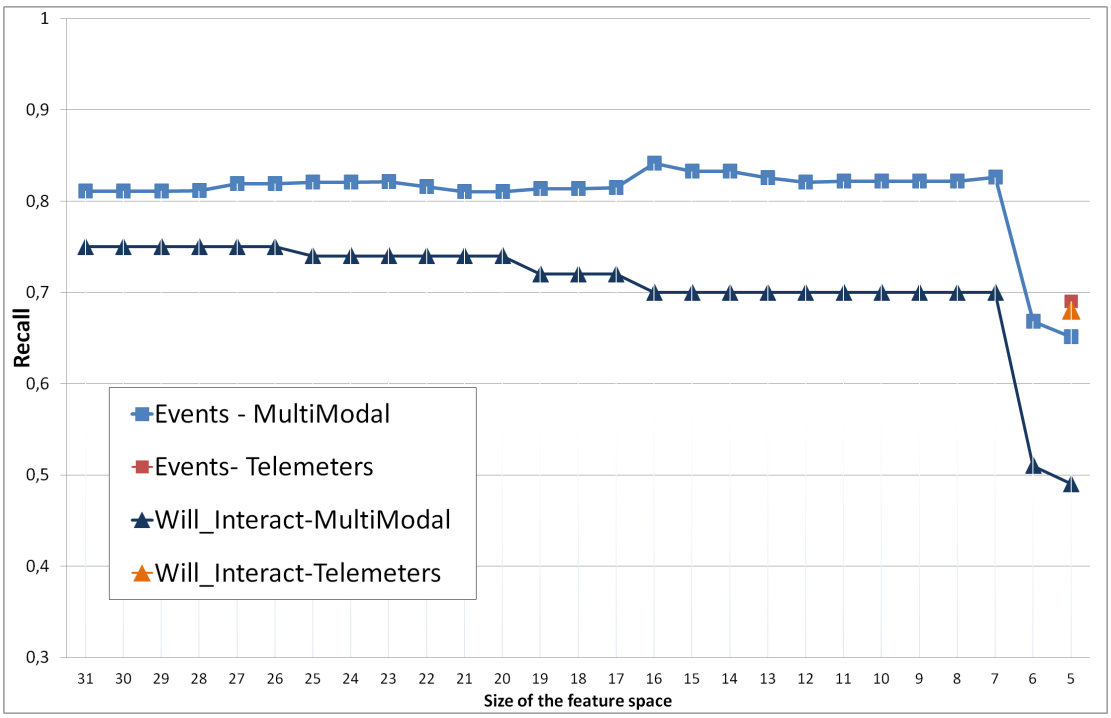


Figure 5.2: Recall evolution with the decreasing number of multi-modal features in comparison with the telemeters for all the events and for the event WILL_INTERACT

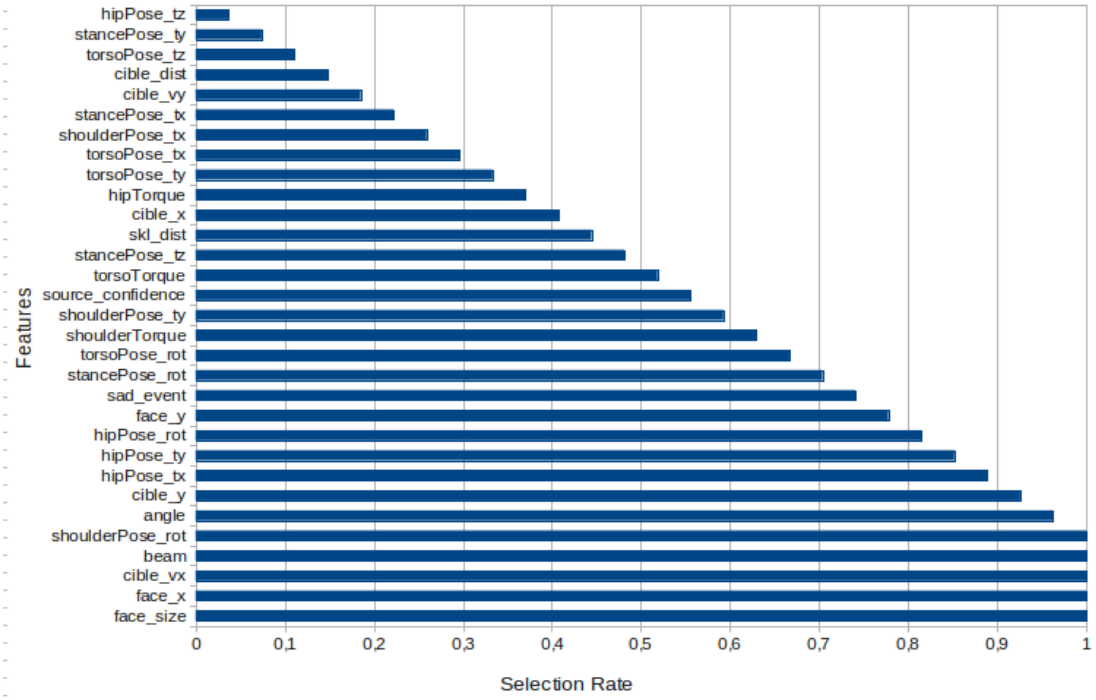


Figure 5.3: Features Selection Rate over MRMR reduction from 32 to 5 features

6 Conclusion

In the context of assistant for leaving robot, psychologists working on the acceptance of robot by elderly and dependent persons in their home environment have pointed the need for more natural and acceptable interaction with the companion robot. Engagement is the first step of interaction. It corresponds to the phase preceding the interaction, when the users shows signals of his wish to interact.

Our goal was to evaluate and measure the clues of engagement of a human involved in a interaction with a robot. Classically detected using the position and the speed of the user, we have shown the limits in term of recall performance of this technique confronting it to realistic scenarios with engagement toward a robot in a homely environment. Indeed, the proximity of a companion robot with the user is not sufficient criteria to predict the engagement. The necessity of testing our hypothesis with real data lead us into a the acquisition of a multi-modal corpus of data.

Building realistic scenarios involving the interaction of participants with the Kompai robot, we have collected various sequences of engagement. Several features were computed over each modality. From the video, we detected faces and used as features the size and position of the face in the images. The skeleton data gave us a clues of the body pose. The sound were used for the sound localization and for the speech activity detection. Telemeters gave us an estimation of the position and the speed of a moving person.

After synchronization and labelling of the features according to the time and the current event (WILL_INTERACT, INTERACT, etc), we merged them into time sequential feature vectors of dimension 33 (the number of features + the event label).

A cross-fold validation allowed us to segment our dataset into a training and testing sets. These subsets where used by two different classifiers, a Neural Networks and a Support Vector Machine classifier. These classifiers trained on multimodal and telemetric dataset gave better recall performances for the multi-modal detection. Indeed, the multi-modality improved significantly the recall of the engagement event.

This last chapter concludes on the accomplish work and gives some of its impacts.

6.1 Lesson learn

As we saw in the results of the evaluation (chapter 5), the multi-modal detection of intention of interaction for a robot companion performance of the multi-modal detection is higher than the one using spatial informations only. We proved that the position and the speed of the human were insufficient to detect his engagement. The MRMR technique helped us to circle some more important features for the engagement detection such as the shoulders orientation.

The size of the dataset constrained the choices of learning technique that we could apply of these data. Indeed, a trial of temporal model Hidden Markov Models has been tempted but did not succeed on account of the lack of prior transitions between the states. The high correlation between the features also made the classification more difficult. Nevertheless, the results on the detection of interaction and intention of interaction classes are improved by the multi-modality.

6.2 Impact of this research

This problem of multi-modal integration for intention detection is a complex problem. This work give a lead to solve this problem and we propose to enrich the corpus with more real life scenarios in order to improve the results. Especially, an enrichment of the dataset will allow us to integrate a temporal aspect in the recognition and to build multi-modal Dynamic Bayesian Networks or a Hidden Markov Models to classify the temporal events that represent our classes.

This master's project is set in the frame of the PRAMAD project for robot assistance to elderly and dependent people. The PRIMA team will soon enrich the dataset by recording data with another robot in Broca Hospital in Paris, partner of the PRAMAD project. A similar approach will be used and the detection of engagement technique that we proposed will be tested in real conditions.

Using a similar method of multi-modal feature fusion the interaction loop between the robot and the human can be finalize by using the engagement detection in order to make the robot detect the availability and solicit the interaction himself when adequate.

This work is one more step in multi-modality for social signal processing applied to human-robot interaction. Indeed, it is felt that the multi-modality can be of a very useful help to decode and recognize affect signals and hence to improve the relationship human-robot.

Bibliography

[1] M. Reasearch, "Kinect for windows programming guide," 2012.

[2] A. Vinciarelli, M. Pantic, and H. Boulard, "Social Signal Processing : Survey of an Emerging Domain," no. November 2008, 1920.

[3] R. Mead, A. Atrash, and M. J. Matarić, "Proxemic Feature Recognition for Interactive Robots : Automating Metrics from the Social Sciences," pp. 52–61, 2011.

[4] M. Pantic, N. Sebe, J. F. Cohn, and T. Huang, "Affective multimodal human-computer interaction," *Proceedings of the 13th annual ACM international conference on Multimedia - MULTIMEDIA '05*, p. 669, 2005.

[5] Robosoft, "Kompai r and d," 2011.

[6] R. W. Picardl, *Affective Computing*. International Series in Expreimental Social Psychology, The MIT Press, 2005.

[7] D. Vernon, C. Hofsten, and L. Fadiga, *A Roadmap for Cognitive Development in Humanoid Robots*, vol. 11 of *Cognitive Systems Monographs*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011.

[8] L. Wang, P.-L. P. Rau, V. Evers, B. K. Robinson, and P. Hinds, "When in Rome: the role of culture & context in adherence to robot recommendations," *ACM*, pp. 359–366, Mar. 2010.

[9] J. Glasnapp and O. Brdiczka, "A Human-Centered Model for Detecting Technology Engagement," *Human-Computer Interaction*, vol. LNCS 5612, pp. 621–630, 2009.

[10] S. Koo, D.-s. Kwon, and E. K. Filtering, "Recognizing Human Intentional Actions from the Relative Movements between Human and Robot," *Nonlinear Dynamics*, pp. 939–944, 2009.

[11] S. H. Kaminski, "Communication models," 2002.

[12] SSPNet, "Social signal porcessing network," 2012.

[13] M. Pantic, R. Cowie, F. DErrico, D. Heylen, M. Mehu, C. Pelachaud, I. Poggi, M. Schroeder, A. Vinciarelli, and Project, "Social Signal Processing: Research Agenda," 2011.

[14] P. Ekman and W. V. Friesen, “The Repertoire Of Nonverbal Behavior Categories, Origins, Usage, and Coding,” 1969.

[15] M. Argyle, *Bodily Communication*. Methuen & Co Ltd, 1975.

[16] P. E. Bull, *Posture and Gesture*. International Series in Expreimental Social Psychology, Pergamon Press, 1987.

[17] K. Jokinen, “Nonverbal Feedback in Interactions,” in *Affective Information Processing*, ch. 13 - Nonve, pp. 227–240, 2009.

[18] P. Krauthausen and U. D. Hanebeck, “Situation-Specific Intention Recognition for Human-Robot Cooperation,” in *LNAI*, vol. 6359, pp. 418–425, 2010.

[19] K. a. Tahboub, “Intelligent Human-Machine Interaction Based on Dynamic Bayesian Networks Probabilistic Intention Recognition,” *Journal of Intelligent and Robotic Systems*, vol. 45, pp. 31–52, Mar. 2006.

[20] J. L. Burke, R. R. Murphy, E. Rogers, V. J. Lumelsky, and J. Scholtz, “Final Report for the DARPA / NSF Interdisciplinary Study on Human âĂŞ Robot Interaction,” *IEEE, TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICSâĂŤPART C: APPLICATIONS AND REVIEWS, VOL.*, vol. 34, no. 2, pp. 103–112, 2004.

[21] H. Knight, “Eight Lessons Learned about Non-verbal Interactions through Robot Theater Motivation : Use Theater to Improve Robot Sociability Background : Non-verbal Interaction,” pp. 42–51, 2011.

[22] C. L. Sidner, C. Lee, and N. Lesh, “Engagement Rules for Human-Robot Collaborative Interactions,” 2003.

[23] P. Holthaus, K. Pitsch, and S. Wachsmuth, “How Can I Help?,” *International Journal of Social Robotics*, vol. 3, pp. 383–393, Sept. 2011.

[24] D. McColl, Z. Zhang, and G. Nejat, “Human Body Pose Interpretation and Classification for Social Human-Robot Interaction,” *International Journal of Social Robotics*, pp. 313–332, June 2011.

[25] A. Mehrabian, “Pleasure-Arousal . Dominance : A General Framework for Describing and Measuring Individual Differences in Temperament,” *Learning*, vol. 14, no. 4, pp. 261–292, 1996.

[26] E. A. Schegloff, “Body Torque,” *Social Research*, vol. 65, no. 3, pp. 535–596, 1998.

[27] N. Dael, M. Mortillaro, and K. R. Scherer, “The Body Action and Posture Coding System (BAP): Development and Reliability,” *Journal of Nonverbal Behavior*, Jan. 2012.

[28] M. Pantic and L. J. M. Rothkrantz, “Toward an Affect-Sensitive Multimodal Human âĂŞ Computer Interaction,” *Organization*, vol. 91, no. 9, 2003.

[29] D. Silva and P. North, “Audiovisual Recognition,” pp. 649–654, 2004.

[30] T. Chen and R. A. M. R. Rao, “Audio-Visual Integration in Multimodal Communication,” *English*, vol. 86, no. 5, 1998.

[31] R. Ooko, R. Ishii, and Y. I. Nakano, “Estimating a User âĂŹ s Conversational Engagement Based on Head Pose Information,” pp. 262–268.

[32] J. Maisonnasse, *Estimation des Relations Attentionelles dans un Environnement Intelligent*. PhD thesis, UNIVERSITE JOSEPH FOURIER DE GRENOBLE, 2007.

[33] K. Karpouzis, A. Raouzaiou, and A. Drosopoulos, “Facial Expression and Gesture Analysis for Interaction,” *Group*, pp. 175–200.

[34] J. M. Rehg, K. P. Murphy, and P. W. Fieguth, “Vision-Based Speaker Detection Using Bayesian Networks,” *Pattern Recognition*, no. Cvpr 99, pp. 110–116, 1999.

[35] A. Jaimes and N. Sebe, “Multimodal humanâĂŞcomputer interaction: A survey,” *Computer Vision and Image Understanding*, vol. 108, pp. 116–134, Oct. 2007.

[36] C. Oertel and S. Scherer, “On the use of multimodal cues for the prediction of degrees of involvement in spontaneous conversation,” *of the International Speech*, pp. 3–6, 2011.

[37] F. Lingenfelser, J. Wagner, and E. André, “A Systematic Discussion of Fusion Techniques for Multi-Modal Affect Recognition Tasks,” *Pattern Recognition*, vol. icmi, pp. 19–25, 2011.

[38] D. Jiang, Y. Cui, X. Zhang, and P. Fan, “Audio Visual Emotion Recognition Based on Triple-Stream Dynamic Bayesian Network Models,” *Audio*, pp. 609–618.

[39] H. J, “Modeling physiological processes with dynamic Bayesian networks,” *Faculty of Electrical Engineering, Mathematics, and Computer Science, University of Pittsburgh*, no. August, 2006.

[40] SSPNet, “Social signal porcessing network,” 2010.

[41] M. Mahmoud, T. BaltruËĞ, P. Robinson, and L. Riek, “3D corpus of spontaneous complex mental states,” *Corpus*, 2011.

[42] D. Vaufreydaz, R. Emonet, P. Reignier, and R. P. Vaufreydaz Dominique, Emonet Rémi, "A Lightweight Speech Detection System for Perceptive Environments," *3rd Joint Workshop on Multimodal Interaction and Related Machine Learning Algorithms, Washington : United States*, 2006.

[43] J. Shotton, A. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman, and A. Blake, "Real-time human pose recognition in parts from single depth images," *Cvpr 2011*, pp. 1297–1304, June 2011.

[44] J. Sanghvi, G. Castellano, A. Paiva, and P. W. Mcowan, "Automatic Analysis of Affective Postures and Body Motion to Detect Engagement with a Game Companion Categories and Subject Descriptors," *Analysis*, 2011.

[45] WillowGarage, "Facedetection."

[46] M. Anne, *Intégration de Services Perceptuels dans une Infrastructure de Communication Ambiante*. PhD thesis, Institut National Polytechnique de Grenoble, 2006.

[47] Sklearn, "Support vector machines."

[48] Weka, "Weka 3: Data mining software and toolkit in java."

[49] P. Hanchuan, L. Fuhui, and D. Chris, "Feature Selection Based on Mutual Information : Criteria of Max-Dependency, Max-Relevance and Min-Redundancy," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 8, pp. 1226–1238, 2005.