HAL
open science

Starting engagement detection towards a companion robot using multimodal features

Dominique Vaufreydaz, Wafa Johal, Claudine Combe

To cite this version:

Dominique Vaufreydaz, Wafa Johal, Claudine Combe. Starting engagement detection towards a companion robot using multimodal features. Robotics and Autonomous Systems, 2015, Robotics and Autonomous Systems, pp.25. 10.1016/j.robot.2015.01.004 . hal-01122396

HAL Id: hal-01122396

https://inria.hal.science/hal-01122396v1

Submitted on 12 Mar 2015

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Starting Engagement Detection Toward a Companion Robot Using Multimodal Features*

Dominique Vaufreydaz[1,2], Wafa Johal[2], Claudine Combe[1]
[1]Prima/Inria-LIG, CNRS
[2]University of Grenoble - Alpes, LIG, CNRS

March 12, 2015

Abstract

Recognition of intentions is a subconscious cognitive process vital to human communication. This skill enables anticipation and increases the quality of interactions between humans. Within the context of engagement, non-verbal signals are used to communicate the intention of starting the interaction with a partner. In this paper, we investigated methods to detect these signals in order to allow a robot to know when it is about to be addressed. Originality of our approach resides in taking inspiration from social and cognitive sciences to perform our perception task. We investigate meaningful features, i.e. human readable features, and elicit which of these are important for recognizing someone's intention of starting an interaction. Classically, spatial information like the human position and speed, the human-robot distance are used to detect the engagement. Our approach integrates multimodal features gathered using a companion robot equipped with a Kinect. The evaluation on our corpus collected in spontaneous conditions highlights its robustness and validates the use of such a technique in a real environment. Experimental validation shows that multimodal features set gives better precision and recall than using only spatial and speed features. We also demonstrate that 7 selected features are sufficient to provide a good starting engagement detection score. In our last investigation, we show that among our full 99 features set, the space reduction is not a solved task. This result opens new researches perspectives on multimodal engagement detection.

Keywords: multimodal perception - affective computing - healthcare technologies - companion robots

1 Introduction

Companion robots are entities that are intended to be used as assistants in everyday life, those being personal coach, desktop manager, etc. They could help to come up with tools that can potentially improve quality of life in the long run. Among usual embedded functions, one can find entertainment, video conference, objects grasping, activity monitoring, serious games and frailty evaluation [37, 8, 10, 9]. Companion robots can assist therapy for autism [6]. This paper presents research on companion robots using the Kompai Robot (see Figure 1)

*Author version.

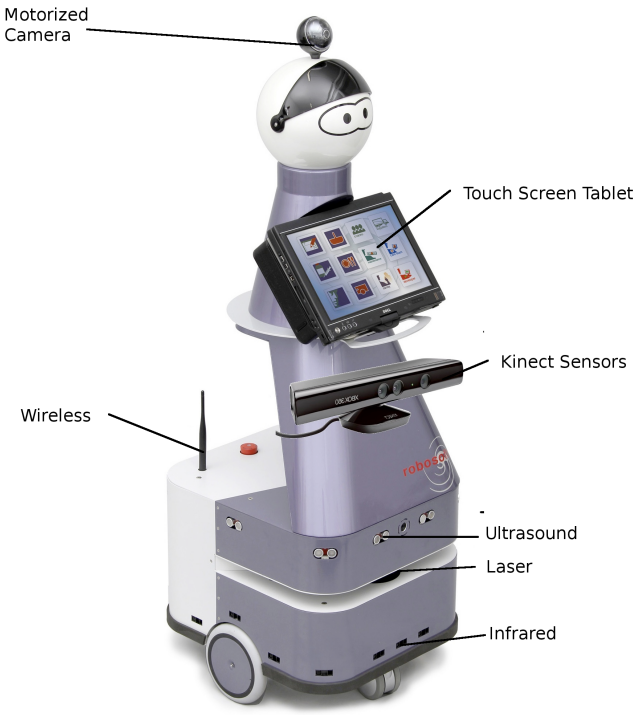

Figure 1: The Kompaï Robot from our partner Robosoft is equipped with a laser range finder, ultrasound and infrared telemeter, a tablet PC and a webcam on top. We added a Kinect for our experiments.

As argued in [20, 29], the primary challenge in building engaging companion robots is to provide social competency in perceiving, reasoning and expressing social and affective aspects of interactions with the human user. Companion robots are aimed to interact with humans in home environments. In order to stay credible, companion robots are expected to behave and react as per predefined manners corresponding to instructions and social signals used by the user.

As speech being multimodal in a face-to-face interaction, non-verbal communication also uses a variety of channels to convey messages. New areas explore techniques for the multimodal aspect of human communication in order to design robots able to read and express communicative signals in a social manner. Non-verbal cues of communications have been well studied for detection of emotions [30, 52, 46, 28]. In this paper, we propose to use these cues for intention recognition, and in particular an intention of interaction.

Recognition of intention is a basic skill acquired by infants early in their development. According to Vernon [45], among other skills, the perception of others' attention 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 [45]. Intention recognition allows the interacting agent to take quick decisions and to respond better to the user's need or state of mind. Some of the non-verbal communication signals are cues to subdued goals and intentions of the humans, and therefore a good way to improve adaptability of the robots' behaviors is by predicting their intentions. A part of human cognition is anticipation, allowing reading intentions and guessing goals in order to react quickly to stimuli. This skill is also very important for turn-taking in interaction.

In neurocognition, the Broca's area, responsible for language comprehension, action recognition & 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.

As humans instinctively detect the intention of someone who wants to ask for way in the street, we are interested in the opening engagement phase of the process during which humans subconciously express their intentions to interact. Our goal is to investigate techniques to detect and recognize signals for non-verbal communication reflecting this intention and in our particular case, the intention of a user to engage an interaction with a robot.

Intention of engagement is a real question, especially when it comes to environments such as the work place or home, where people are not used to interact with robots [48]. Classically, the criterion for a user's intention of engagement is the spatial distance between the user and the communicant interface [11]. Some investigations have improved on this idea by also considering the speed of movement of the user [19]. These studies have chosen to use the relative spatial position of the concerned agents as criteria. The following assumption is made behind this choice: if the user is close to the robot, there is an intention to interact. Using distance and sometimes speed of the human provides with satisfactory results, but for a companion robot in real situations at home, close distance does not necessarily signal a desire for engagement. For instance, many times during the day, one can pass in front of the refrigerator without the wish to open it. Following the same logic, despite the physical distance of the user from the robot, a robot should be able to detect when it is about to be solicited, and anticipate the interaction in order to be more comfortable and socially acceptable.

In this study, we propose a multimodal approach for detecting a starting engagement using a RGB-D sensor mounted on a companion robot. Getting inspiration from social and cognitive sciences, our goal is to select features in order to improve the re-usability in other situations and/or with other sensors. In our approach, the idea here is to get rid of the usual way to do such experiment i.e. putting all available features together, combining them in a more optimized representation and let the training paradigm filter everything. Doing this, we might have good performances, but we may not learn anything about detecting intention of engagement. We will see that less than 10% of our features are crucial for starting engagement detection. In another context, one can make well-founded choices among sensors to reflect this knowledge. It will be more efficient to design a new device or robot knowing which particular features are of importance. This prospective research aim to build a set of meaningful features extracted from multimodal sensors useful for the description, recognition and discrimination of the intention of engagement.

This paper aims to contribute on the following statements :

- There exist subconscious social signals expressed by humans that characterize their will to interact with a robot and these signals are detectable.
- Some features from literature in the social and cognitive sciences are computable on a companion robot (notably Schegloff metrics [36]).
- Multi-modal features will perform better than spatial features to detect this starting of engagement in a home-like environment. A realistic dataset in a home-like environment can help us to validate this hypothesis.
- The set of relevant features for starting of interaction detection can be reduce without loss of performance using a feature space reduction process using the Minimum Redundancy Maximum Relevance (MRMR) method [13] never used in this context.

2 Multimodal Social Signal Processing For Non-Verbal Communication

2.1 Social Signal Processing

A communicative agent does not use only the verbal channel, but many channels to send and receive various messages while interacting [16]: human communication is intrinsically multimodal. To make human-robot communication fluent and acceptable, the robot has to decode these behavioral and non-verbal cues in order to act accordingly. For instance, computer systems and devices able to recognize

agreement or inattention, 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[42, 45]. In the context of people assistance, social features seem to be crucial for the acceptance of a robotic companion in a domestic environment.

Argyle, in his book "Bodily Communication" [1] mentions different signals from different modalities used for non-verbal communication. The considered modalities are facial expressions, gaze, gestures & body movements, posture, contact, spatial behavior, clothing, and vocalizations. This work shows that recording of these modalities allows to recognize the mood of a person.

P. E. Bull [4] follows the idea that communication implies a socially shared signal system or code. Non verbal communication is argued to be intentional or non-intentional. Therefore, it is valuable information that allows to access intentions of the emitter that can be non-voluntarily transmitted. Bull claims the importance of posture and gesture in non verbal communication where these channels have been neglected compared to facial features and speech cues.

2.2 Intentionality in Human-Machine Interaction

The intention cues form a way of communication. As stated before, recognition of human's intentions, goals and actions is important in the improvement of non-verbal human-robot cooperation. [21] defines intention recognition as the process of estimating the force driving humans' actions based on noisy observations of human' interaction with their environment. Tahboub in [43] sees intention recognition as a substitution or a complement to reliable and extensive communication that is a prerequisite for coordination and cooperation. Indeed, in order to have a smooth interaction, intention recognition is essential. The DARPA/NSF final report on Human-Robot Interaction [5] recommends to improve the models of human-robot relationship and in particular to work on the intentionality issue.

In [19], it is proposed to recognize intentional actions using relative movements of a human towards a robot. An IR sensor embedded on the robot is monitored to track and estimate the velocity of a person. They then infer intentional actions such as approaches and departs using Hidden Markov Models (HMM) and position dependent models. Other related studies have defined human-robot proxemics in order to adjust inter-personal distance [47]. This work seems not enough to estimate engagement. On one hand, one can slow down near the robot without wishing to interact. On the other hand, someone passing swiftly by close to the robot might want to interact with it hastily.

The relative position and speed are not the only features that should be used to estimate intentionality. Multimodal fusion and usage of postural information have given good results in the measurement of quality of human-robot interaction and engagement of the user into the interaction [6, 34]. We aim to detect intention of interaction using also postural information and not only proxemics features.

In his study, Knight [18] points towards the importance for a robot to convey and hence to detect intentionality. It helps to clarify current activity and to anticipate goals. Learning from the engagement of humans, the robot should be able to anticipate the interaction and also to learn adequate moment when the robot itself can engage an interaction. In [39], 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. Sidner et al. propose a model in three steps: (1) initiation of interaction, (2) sustainance of interaction, (3) disengagement. As presented in 3.3, we will see that our classification is based on this model.

3 Corpus for Engagement with a Companion Robot

A part of the work accomplished was to build a multimodal dataset including interactions with a companion robot equipped with a laser telemeter and a Kinect device. In this context, we focus on working with consumer devices and in a natural and non intrusive manner. Even though the tendency is the increasing usage of physiological sensors, such as R. Picard's pulse bracelet called Cardiocam, physiological signals still remain an invasive and relatively expensive option for users, for them to be released widely. The

physiological modality is not considered in this work, yet it might be enriching to include it in a future work that uses, for instance, contact-free heart rate measurements [31].

In order to validate our hypothesis in the context of interactions with a robot companion, the considered sensors are the ones commonly found on such robots: microphones, video sensors, depth sensors and telemeters. There exist available datasets in the field of social signals processing dealing with non-verbal communication which use multiple sensors. These datasets for emotion recognition are unfortunately more often built for face-to-face interaction where people sit and interact only with speech. The SSPNet association released the SEMAINE-DB dataset [41] where several persons were recorded in a face-to-face speech interaction. This database is suitable for a desktop environment that involves an interaction with a virtual commmunication agent. It is not well suited for human-robot interaction; especially as body cues in social signals are more diverse than facial expressions and speech characteristics. There exist other datasets using the Kinect sensor; such as Cam3D dataset centred on facial and hand movement associated with audio recording [22], or the LIRIS Human activities dataset [51] associated with human activity monitoring task. However, the proposition of a robot centred dataset for multi-modal social signal processing has not been made yet.

Looking at limitations of the existing multimodal datasets, the sensors equipping the Kompaï robot have been used to record a new robot view-point dataset where the users are interacting with the robot while standing. The scenarios included in this dataset will be presented in section 3.4.

3.1 Realistic Dataset

R. Picard in [30] states five variables that may affect data collection. (1) The first factor is the spontaneity of the behavior. The emotion can either be elicited by a stimulus or acted. (2) Another influence can come from the environment of the recording and the question here is: are the expressions of the participant similar in a lab setting and in a real-life situation? (3) Next question to be considered when recording affective data is: should the focus be on the expression or on the internal feelings of the participant? (4) The participant's awareness about the fact that he's being recorded. Indeed, what is the influence of open-recording in comparison with hidden recording on the recorded data? (5) Finally, should the participant be informed of the purpose of the experiment?

Regarding this research matter, (1) the engagement is relatively spontaneous, because the participants didn't act the interaction but were asked to interact whenever they wanted to with the robot. (2) The recording is made in a *living lab* environment, similar to a flat. The participants have no prior experience of this environment. This can create some fluctuations in their behavior. (3) We wanted to record intentionality of interaction, hence we focused on expression of social cues rather than to do a subjective evaluation. (4-5)We chose to not tell the participants that we were interested in the social cues of intention of interaction in order to collect more natural data.

3.2 Experimental implementation

The experimentation space is presented in figure 2. The apartment is divided into 3 areas: a living-room, a kitchen space and an empty space. To test our assumption that spatial information is not enough to detect an intention of interaction, furniture is placed so that participants will need to pass near the robot each time they want to go from one side to another of the experimentation room, even if they do not want to interact with the robot. This choice is an adverse condition as it leads us to distinguish someone passing close to the robot with or without intention of interaction.

In our recording, the robot is immobile. All features are robot centered but can also be computed with a mobile robot. The interaction in this dataset consists of playing a "tap the mole" game on the mounted tablet PC on the robot. Our hypothesis was that the interaction with a companion robot is also preceded by a pre-interaction phase (see section 3.3) where the participant shows some subconscious social signals of its intention of interaction. We also assume that these cues are detectable with the sensors that are equipped in our enhanced version of the companion robot (see figure 1).

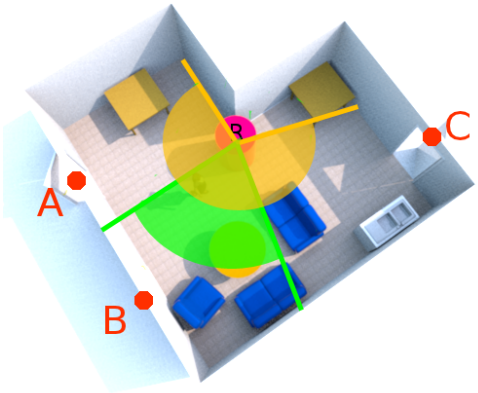

Figure 2: Home-like environment for our experimentation. The area has an L shape and 3 access doors (A, B and C). It is organized around 3 spaces: a living-room (near door B), a kitchen space (near C) an empty space (near A). The robot is place in the center of the area (purple cylinder). The view from the Kinect is depicted in green and the telemeter field of view in orange.

3.3 Steps of the interaction process

Sidner et al. in [39] proposed a model to describe the process of interaction in three steps: (1) initiation of interaction, (2) maintenance of interaction and (3) disengagement. Our work follows this approach by modeling these events as (illustrated in section A.1): (1) WILL_INTERACT, (2) INTERACT, (3) LEAVE_INTERACT. We added two more classes. The SOMEONE_AROUND event is when someone is detected in the room but with no wish of interacting with the robot. When nobody is in the room, it corresponds to the NO_ONE event.

3.4 Scenarios

The data is recorded in two different scenarios performed several times by one or several participants in a home-like environment where the Kompaï robot is present. The first one is dedicated to mono-user experiment. Only one user will be in the room at a time. The multi-users scenario addresses a more adverse condition. Three persons are already in the room and interact with each other. The idea behind this scenario is to check if we can detect starting engagement among social interaction between participant.

Each participant was given randomly one or several actions to perform in the room. As said, the room is similar to a small flat (Figure 2). It was asked to the participant to enter the room by different doors, perform some realistic actions and to go out. One action is to interact with the robot. The other actions were going across the room, walking, sitting, playing cards or pouring water from the sink.

3.4.1 Scenario 1: Passing By

In this scenario, each participant is asked to go through the room by different doors (A), (B) or (C). At this point, the given instructions did not mention the robot's presence in the living lab. Participants were not aware that they will interact later with the robot. After some crossings, the participant was invited to play the game on the robot's tablet. The Figure 3 shows the setting of this scenario.

3.4.2 Scenario 2: Playing cards together

In this second scenario, 3 or 4 persons were asked to enter the room and start playing cards in the living-room area. A telephone placed in the room was used to ask one of the participants to execute an action (interacting with the robot, or using the sink for instance). Once the participant was asked to perform a task, he could do it when he wanted to. The participants could sit wherever they wanted in

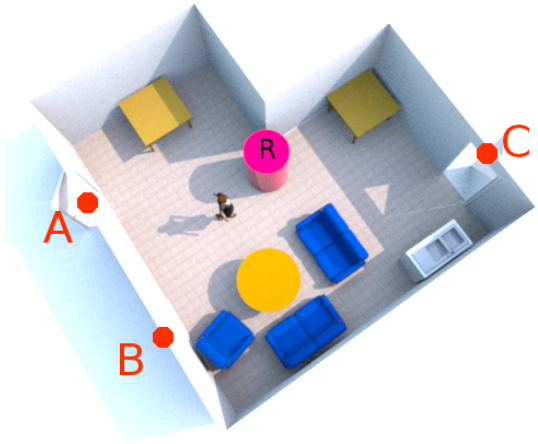

Figure 3: Scenario 1, Passing by

Figure 4: Scenario 2, Playing cards together

Figure 5: Perception from the robot point of view with a coming user toward a robot. Laser Telemeter (red lines), foot (blue spots) and pedestrian (green ellipse) information are depicted on left picture. In the middle, one can find RGB view from the Kinect with face detection (green circle). Right, the depth view with user (blue) and skeleton (yellow) tracking are drawn (note that with the first Kinect version, *Kinect for XBox*, there is a little shift between the RGB and depth views). Acoustic and other body features are computed on these data (see section 4).

the living-room area. The figure 4 shows this scenario when a new participant is entering in the room while two participants are already sitting.

4 Features Extraction

In order to characterize the engagement, features were extracted from the corpus previously introduced and then synchronized with a unique time scale. Our Kompaï robot, loaned by our partner Robosoft[1] (see figure 1), is composed of a 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 pair of microphones, a motorized web camera and a speaker device. We added a Kinect sensor to the robot. Novelty of this work is not the adjunction of the RGB-D device but the synchronous usage of information from all sensors to compute a multimodal feature set. The current version of the system is online and computes on the fly all features on the Kompaï.

[1] `http://www.robosoft.com/`

Feature extraction algorithms present a fair amount of noise in general and an interest of using multimodality is to be able to compensate one modality with another. The full feature set gathered and computed on our corpus is composed of 99 features. Then, a feature selection is made driven by social and cognitive science research on non-verbal communication cues depicted in section 2, the availability of sensors on our Kompaï robot and the performance of algorithms in our experimental conditions. We let aside important cues that might improve our results but are not usable in our context. For example, gaze direction and facial emotion recognition can not be computed; hand state and gestures are not reliable for instance. Raw (x, z, y, *confidence*) tuples for skeleton joints permit us to compute more interesting body features.

This section presents different feature extraction techniques used in our experiment. We chose to investigate a selection of 32 features including spatial information (spatial subset), body pose and video face detection (body subset), speech activity detection and sound localization (acoustic subset) in order to model the intention of engagement. We detail these subsets in the following subsections. These features are computed on several raw data channels: laser telemeter data coming from the robot, rgb video (section 4.3.3), depth view (section 4.3) and audio channels from the Kinect (see figure 5). The synchronization and labeling methods are then explained in section 4.4.

4.1 Spacial features

Figure 6: Spatial features: foot and pedestrians tracking using robot telemeter. Blues points are foot, ellipse represents tracked pedestrian. The green lines represent, in the robot frame, the Kinect field of view. On this figure, there is one tracked pedestrian. The pedestrian is out of the camera view.

Proxemic features are classically used to describe role, attention and interaction, and in particular to determine the intention of interaction. The tracking of the human trajectory can be done through visual based models or using laser telemeters. Telemeters provide planar information of the environment while covering a wider range angle than standard video camera with a good precision. The Kompaï robot is equipped with a single-row laser-range scanner at 20 cm above the ground. For pedestrian tracking, it is more likely that we can detect shins.

Classical proxemic features are the relative position of the individual to the robot and his speed. For a successful collaboration, the distance between the robot and the human should be optimum and the speed controlled [19]. It is important to know about distance so the person interacting with the robot does not feel uncomfortable [12]. [53] proposed a system for tracking pedestrians using multiple single row

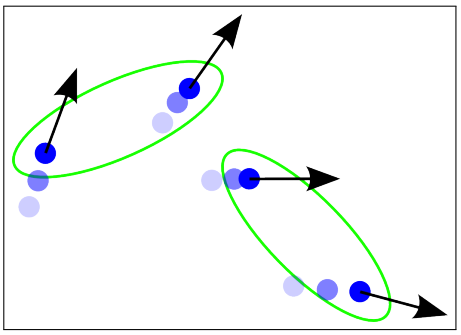

Figure 7: Feet grouping if the pair of foot satisfies a constant space between legs through time.

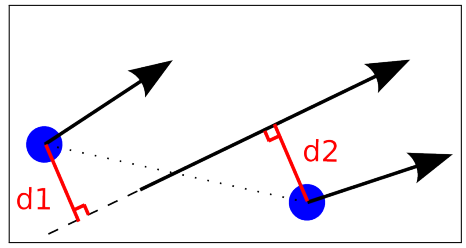

Figure 8: Space between legs is given by the sum of d1 and d2, the distances between feet and the main direction vector.

laser scanners. Their pedestrian's walking model is described and used to accurately track pedestrian feet according to their swinging phase. In our study, since a single scanner is used, foot occlusion is frequent as soon as a foot goes behind the other from the telemeter point of view. Therefore, it is difficult to predict swinging phases because of sparse data. Our human detection and tracking process is done through a feet-pairing process where a human is represented either by its both feet or one single foot when the other one is temporary hidden.

4.1.1 Feet detection

The laser sensors that equip the Kompaï robot give the distance values over 270 degrees every 80ms. An adaptive background subtraction on the telemeters values is used to detect moving objects in the room. Moving objects are candidates for being feet. The distance and the angle associated to the detected moving object are used to compute the positions x and y into the robot reference frame (see fig. 6).

4.1.2 Speed Estimation

A Kalman Filter is used on the set of moving objects detected by the laser to compute the speed and the acceleration 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 a moving object) and to estimate dynamics such as the position and speed in two dimensions ($cible_dx, cible_dy$). The implementation of the Kalman Filter over the telemeters moving object data has been made using the OpenCV library. A direction vector is extracted from each foot tracker. It will be used for feet pairing in pedestrian tracking.

4.1.3 Pedestrian tracking

Grouping feet simply according to the distance between feet is not sufficient enough since pedestrians have different step length. Furthermore people standing side by side could be misidentified. Looking at the space between legs rather than between feet leads to a more robust parameter: even if the step length varies for the same pedestrian (standing, walking, running) the distance between legs along a direction vector remains constant during a natural walk because of the geometric properties of the human skeleton. Our process consists in pairing tracked feet that match a particular model, using a 2 stage filtering. First, for each frame, feet that are less than one meter apart are paired up together forming a potential pedestrian. This one meter threshold was empirically set to quickly exclude impossible pedestrians. Then, candidates are evaluated and an actual pedestrian is revealed if it satisfies that the space between legs through a short frame sequence is relatively constant ($\sim$30 cm) as shown in figure 7.

The space between legs is computed as the sum of the distance between each foot and their projection on the main direction vector, which is the sum of two single foot direction vectors, figure 8. Pedestrian targets can be initialized as soon as they are in movement. One of the feet can then be hidden without causing the loss of padestrian's localisation, it will be paired up when the foot appears again. If both feet disappear, then the pedestrian tracker is lost and deleted.

Our tracking process is capable of tracking multiple walking pedestrians with frequent occluded feet from a single range laser telemeter.

Features Name	**Sensor**	**Frequency**
Positions ($cible_x, cible_y$)	laser range finder	12.5Hz
Speed ($cible_dx, cible_dy$)	laser range finder	12.5Hz
Distance ($cible_dist$)	laser range finder	12.5Hz

Table 1: Features from the Space subset

Every 80 ms, we have the *Number of pedestrians* around the companion robot and for each pedestrian, we get an *id*, *cible_x* and *cible_y* position in the reference frame of the robot, his distance to the robot *cible_dist* and *cible_dx*, *cible_dy* the speed of the pedestrian in x and y axis (see Table 1 above).

4.2 Acoustic features

Pantic in [28] and [46] list some features from the audio signal that can be used to spot basic emotions 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. Moreover, speech is an important information source for social glue with a companion robot [2]. Considering the recognition of the starting engagement in an interaction, only few papers in the literature use audio features in a multimodal frame. [27] proposes an engagement estimator using head pose associated to audio features in a face-to face conversational agent sitting interaction. Some articles invoke interest of sound localization in attention or focus estimation [23].

Features Name	**Sensor**	**Frequency**
Speech Activity (sad_event)	Kinect's Microphones	100Hz
Source localization ($beam, angle, confidence$)	Kinect's Microphones	8Hz

Table 2: Features from the Acoustic subset

The microphone array embedded in the Kinect sensor is a four-element linear microphone array processing acoustic echo cancellation and noise suppression. Using this audio stream, we can compute Speech Activity Detection (SAD) [44], which is indicative of the parts of the acoustic signal representing speech. The SAD labels the audio stream every 10 ms. The source localization outputs the stimulated beam (rough estimation) and the source position (more accurate angle) associated with a confidence. The frame rate of the acoustic localizer is 8Hz.

4.3 Body features

The Skeleton tracking of the Kinect sensor allows real time pose and gesture recognition. Our system outputs at depth camera frame rate the *number of skeleton*, and for each skeleton an *id* and 60 features giving x, z, y and *confidence* for each joint.

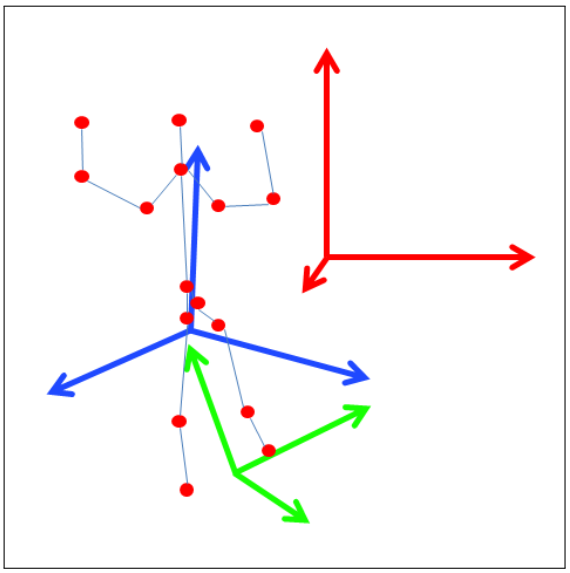

Figure 9: Stance (green) pose, hip (blue) pose and torque. Body pose features computed from the skeleton information of the Kinect sensor (positioned in red)

4.3.1 Body pose

As expressed previously, body pose features give clues on intention of interaction. These features can measure the level of engagement of a user into a task as in [35], proposing a measure of the Body Lean Angle. Psychologists have proposed many models to describe body pose metrics and their associated meaning. An overview of these metrics can be found in [24]. In [14], the authors propose a spacial model coupling attention estimation and distance metrics for a receptionist robot to infer the intentions of the human. Their results are promising, but the approach is limited to face-to-face interaction. [15] studied on spatial relationships in human robot interaction and concluded that human-human proxemic measures and social arrangement such as Hall's interpersonal distance system are not enough to achieve socially appropriate robot behavior. Psychologists such as Hall [12], Mehrabian [25] Schegloff [36] 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 new devices like the Kinect Sensor and other real-time 3D pose reconstruction systems, we are now able to evaluate the pose of a person.

The body features used in our experiments are based on Schegloff metrics presented in [24, 36] and computed from the Kinect skeletons. These features aim to depict the body pose of the individual. The accent is posed on the stance, the hips, the torso and the shoulders' position and orientation relatively to each other. We obtain 19 features containing Schegloff's metrics at depth frame rate.

4.3.2 Distance

What is interesting about body features is that they depict the orientation of the bodypart relatively to the Kinect sensor placed on the robot. A *skeleton distance* associated to the skeleton position is computed using the average z-value of several joints of the skeleton.

4.3.3 Face detection

In terms of affect & emotion detection and speech recognition, a lot of studies have published results with a combination of face and audio features [33, 40, 7, 17]. Within engagement, the orientation of the head and the gaze seem to be crucial. As shown in [32], a speaker can be detected more easily with the combination of different features as a mouth sensor. Face detection is already a first cue of interaction,

and orientation of the face toward the robot is a reliable sign of attention. Gaze tracking can give a better estimation of user' attention, but performance in uncontrolled real-life condition (small faces in video stream or untrained gaze angle for instance) are not good enough.

From the video extracted from the RGB stream, we propose to focus on face detection. We use a trained machine learning system using Haarcascasdes method. The training is provided by the OpenCV library [50]. The gathered features are the position of face(s) in the pixel reference frame, the {*0,0*} point is the center of the image. For each detected face, we have x, y and *face size*.

Features Name[2]	**Sensor**	**Frequency**
Stance ($*Pose_x, *Pose_y, *Pose_z$, $*Pose_rot$) for feet, hips, torso and shoulders	Kinect's Skeleton	30Hz
Relative torque angle ($*Torque$) for hips, torso, shoulders	Kinect's Skeleton	30Hz
Skeleton distance (skl_dist)	Kinect's Skeleton	30Hz
Face ($face_x, face_y, face_size$)	RGB stream	30Hz

Table 3: Features from the Body pose subset.

4.4 Fusion, Synchronization and labeling of features

At this point, we have a selection of 32 features: pedestrian information (x, y, *speed_x*, *speed_y* and *distance* to robot), Schegloff metrics (computed from the skeleton see 4.3.1), face detection, speech activity detection and sound localization. This section details how we deal with sparse features, features synchronization and corpus labeling.

4.4.1 Sparse features

Multimodality has one major drawback. Space coverage is not the same for all sensors: the Kinect has a 60 degrees field of view, the laser telemeter 270 degrees, etc. Moreover, we do not have every feature all the time. Whereas video, depth and laser telemeter data, face detection, sound classification, skeleton and pedestrian tracking are not available all the time. One way to cope with these sparse data is to train several classifiers with all possible combinations for available features and to select the adequate one at the right moment. The problem with this approach lies in reducing the amount of data for training for each subtype of classifiers. Another way to solve this problem is to use specific neutral values for unavailable features. For example, when there is no pedestrian, we can set all pedestrian features (position, speed and acceleration) to 0. This set of values is considered neutral as it is impossible to find them in observed data. In these experiments, as we did not have enough data to train each subtype of classifiers, we chose the second method.

4.4.2 Features synchronization

We needed to synchronize the monitored data from the different modalities. Data collected through the Kinect sensor such as the skeletons' positions, the video and the depth are tagged with a time relative to the Kinect sensor's initialization. The laser data are labeled with an absolute time stamp thanks to the real-time micro-controller of the Kompaï robot. The telemeters' input is the steadiest one at a fixed 80 ms period, hence it is used as synchronization frame rate at 12.5HZ. The short time delay between frames prevents us to interpolate and allows to elicit the last value of each feature as the current value.

[2]Suffixes are presented with a "$*$" character. For example, we compute *shoulderPose_x*, *shoulderPose_y*, *shoulderPose_z*, *shoulderPose_rot* and *shoulderTorque*.

4.4.3 Corpus labeling

The labeling of the dataset with the 5 classes (WILL_INTERACT, INTERACT, LEAVE_INTERACT, NO_ONE, SOMEONE_AROUND) is semi-automatic. The timestamped notes of the experimenter contain start time and end time of all events (participant p is entering the room, a specific action was asked, etc.). This annotation serves as first segmentation input. The first labels are then made automatically using both the tablet touching information and the available features. The INTERACT time-interval is labeled from first touch of the tablet until the last click. A WILL_INTERACT event starts when someone is coming to the Kompaï robot before an INTERACT event. If someone is moving or sitting, and after decides to come to the robot, only the direct path to the robot is labeled as WILL_INTERACT. The LEAVE_INTERACT labeling is made like the WILL_INTERACT even and corresponds to all direct paths after leaving the interaction. The NO_ONE and SOMEONE_AROUND events, correspond to the rest of the time, respectively, there is no one in the room and when a person is present, but there will be no interaction.

All labels have been reviewed by a human expert looking at the video recordings. No problems were reported as a result.

4.5 The dataset in numbers

In our dataset, each frame corresponds to 80 ms, provides a full feature set and has a unique label. The total number of recorded frames is 158200[3]. The number of synchronous frames for each event is not equal. The Figure 10 shows the data distribution of each event.

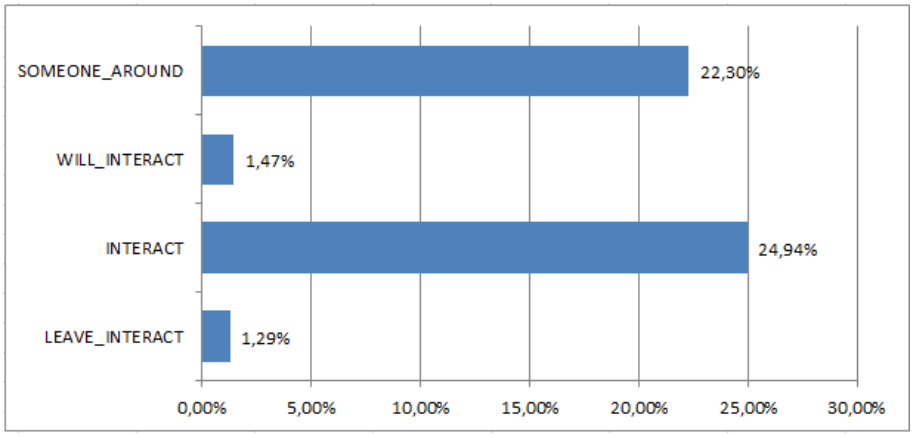

Figure 10: Percentage of each class in the dataset.

In real life, individuals do not express social signals the same way. A certain variability was introduced in the pool of 19 participants. They were from 20 to 35 years old, almost 50% male/50% female, students, administrative assistants and researchers. Voices, clothing styles (colors, trousers or skirts, etc.) and statures vary to challenge perception algorithms. 15 participants did 1 or 2 interactions from 2 to 10 minutes according to their will, 9 were recorded both in mono-user and in multi-users scenarios.

In total, the corpus includes 29 interactions with the robot, made by 15 different participants. The total size of the uncompressed data set is around 300 GB. One can find samples of the corpus in A.1.

5 Multimodal detection of engagement

In this work, we first choose to test all the modalities that can help us to detect intention of interaction. Then, a selection can be made among the most relevant multi-modal features (section 5.2.3). The evaluation focuses on comparing the detection of the intention of interaction by using multimodality

[3]The total recording time is 3:30:56.

Telemeters condition

Class	Precision	Recall
NO_ONE	0,95	1,00
WILL_INTERACT	0,91	0,77
INTERACT	0,77	0,96
LEAVE_INTERACT	0,00	0,00
SOMEONE_AROUND	0,75	0,35

Multimodal condition

Class	Precision	Recall
NO_ONE	0,95	1,00
WILL_INTERACT	0,90	0,87
INTERACT	0,84	0,95
LEAVE_INTERACT	0,21	0,01
SOMEONE_AROUND	0,76	0,41

Table 4: Results for Neural-Network 5-classes classification using Weka. Left table presents results for the telemeter condition. Right table for the multimodal condition.

versus simple spatial information. We want to confirm that these state-of-the-art approaches based on spatial features only are not enough in home-environment with furniture. The Scipy library through Sklearn [38] and the Weka toolbox [49] were used for the classification. The techniques used for the classification are the Multi-class Support Vector Machine (SVM) from Sklearn and the Artificial Neural Network (ANN) technique from Weka.

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 a training set and a test set. A way to randomize this splitting is the k-cross folding process. In this method, the dataset is partitioned in k subset. One subset 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 against others subsets. K-cross validation allows to ensure that the splitting is quite random. Since the events (interaction phases) are not equally probable and temporally related, we used a *stratified* k-fold-cross validation that keeps the same proportion of the different classes in the splitting process.

For our experiment, using $k = 10$, the train and test sets are composed respectively of 140292 and 15587 frames[4].

5.2 First classification experiment

We chose to use two kinds of classification in this experiment. Even if many other techniques could have been applied, we decided to focus on Neural Network and Support Vector Machine (SVM) (sections 5.2.1 and 5.2.2). For this two techniques, we built and tested two classifiers one for the multimodal condition (including 32 features) and one for laser telemeter only condition (a subset of the multimodal dataset including spatial information only), see [3]. In 5.2.3, we try to determinate if some features are more relevant for our task.

5.2.1 Neural Network

The Artificial Neural Network is a graphical layered model commonly used to infer model from observation. In our case, we suppose that our features set can characterize the starting of engagement. ANN is a good classifier to build prospective detection especially with large features vector. The test results of the ANN classification are presented in left table in 4 for the telemeter, and the right table for the multimodal dataset. Notably, one can see that that LEAVE_INTERACT was not classified in the telemeter condition. The INTERACT precision increased in multimodal condition combined with a small loss in recall. Concerning the WILL_INTERACT class, the system returns more relevant events using multimodality (higher recall score) even if its precision decreased. In multimodal condition, the precision is improved for most of classes. The Neural Network classifier gives always better recall rate in this condition.

[4]Note that the *stratified* splitting process let aside 2321 frames.

Telemeters condition

Class	Precision	Recall
NO_ONE	0,68	1,00
WILL_INTERACT	0,80	0,68
INTERACT	0,00	0,00
LEAVE_INTERACT	0,00	0,00
SOMEONE_AROUND	0,76	0,01

Multimodal condition

Class	Precision	Recall
NO_ONE	0,92	0,88
WILL_INTERACT	0,92	0,71
INTERACT	0,54	0,77
LEAVE_INTERACT	0,04	0,10
SOMEONE_AROUNDd	0,52	0,29

Table 5: Results for SVM 5-classes classification using Sklearn. Left table presents results for the telemeter condition. Right table for the multimodal condition.

For the intention of engagement detection, in a practical point of view, the accent has to be put on the good performance in term of recall associated to a low false-positive rate. Using Neural Network in multimodal condition seems to fulfill this requirement.

5.2.2 Multi-Class Support Vector Machine

The results of the 5-classes classification using support vector machine for the multimodal condition are presented on right table in 5, on left table for the telemeter condition. Analyzing these tables, one can see that the False-Positive rate is higher in the telemeter condition. The INTERACT class is not classified at all. The precision and recall scores for WILL_INTERACT class are improved by the multimodality. The aim of our work was especially to decrease the rate of misclassifying an event as WILL_INTERACT, hence the system has fewer chances to predict an interaction when there is one user with no intention of interaction. In the case of an SVM classifier, multimodality is thus interesting for this purpose.

Scores for INTERACT, LEAVE_INTERACT and SOMEONE_AROUND classes are of interest. In multimodal condition, the SOMEONE_AROUND scores drop while INTERACT and LEAVE_INTERACT are better classified. This fact is a first clue of the closeness of our classes in the feature space. The section 5.2.4 will discuss this topic.

5.2.3 Minimum Redundancy Maximum Relevance experiment

A dimensional reduction of the features space was made using the Principal Component Analysis (PCA) and the Linear Discriminant Analysis (LDA) using the Sklearn tool-kit. The results were not conclusive, the dimensionality reduction gave strictly the same performance during the classification where we were expecting an improvement.

The Minimum Redundancy Maximum Relevance [13] (MRMR) technique was performed in order to highlight the best features for our detection system. This dimensionality reduction technique has the advantage of giving the more relevant features instead of building new features from the observed ones. Using mutual information, correlation and t-test/F-test metrics, the MRMR algorithm selects a feature subset maximizing dissimilarity of features and statistical characterization of the classification. It could allow eventually to discard less relevant features in order to optimize the detection of engagement process.

In order to evaluate the relevant features for the multimodal detection of intention of interaction, we used a MRMR dimensionality reduction from a vector of 32 features before performing an SVM learning. The Figure 11 shows the feature reduction's impact on the precision. The precision drops at the 6-features reduction. From the 32-features till 7-features along the feature space reduction, the precision remains pretty stable. These results confirm that there are many redundancies in the 32-feature space. Some of these features seems to be fundamental for a better detection with a higher precision than the telemeters' one. Equivalent conclusions can be made on the Figure 12 regarding the recall performances.

The first remark on these results is that the seven 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 Kinect view. The *beam* and the *angle* are the sound localization features from the microphone array. The telemeter information are considered as relevant, with the high selection rate of the speed *speed_x*

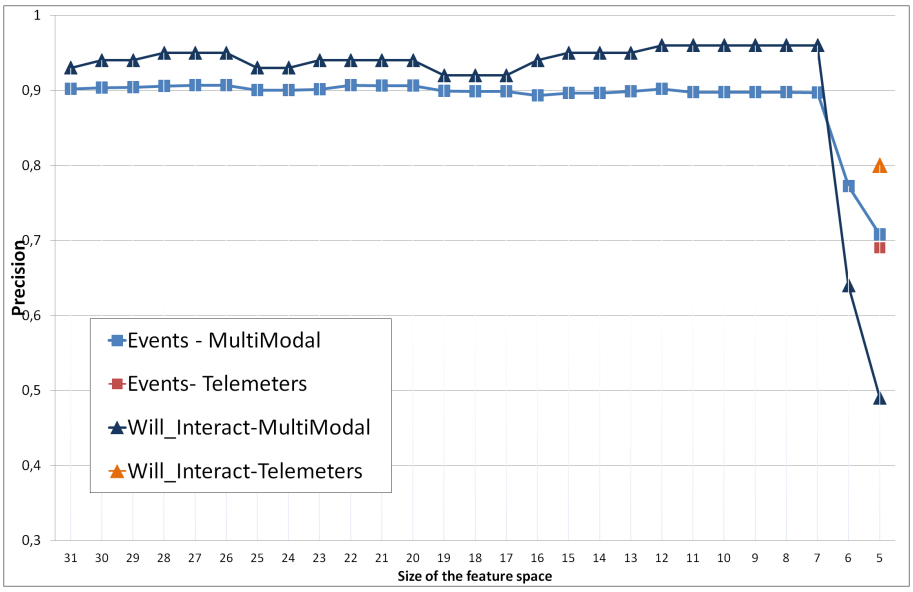

Figure 11: Precision evolution with the decreasing number of multimodal features in comparison with the telemeter condition for all events and for the WILL_INTERACT event

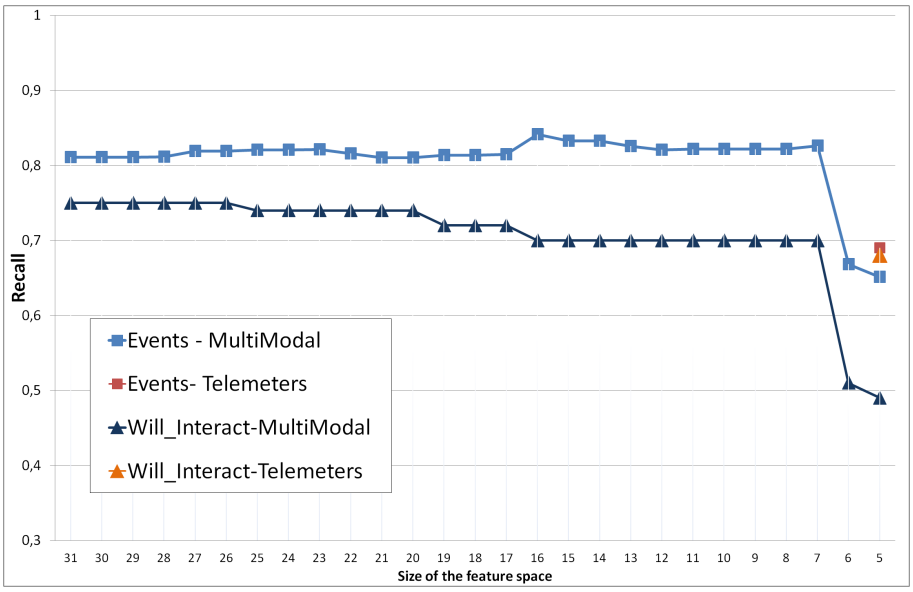

Figure 12: Recall evolution with the decreasing number of multimodal features in comparison with the telemeter condition for all events and for the WILL_INTERACT event

and y position. The fact that these spatial features are selected among the most relevant ones in our multimodal set is not surprising. They were part of previous state-of-the-art researches. Last, the *shoulder pose rotation* corresponds to the relative orientation of the shoulder in the body, and is extracted from the skeleton information.

5.2.4 Discussion about first experiment

The previously presented results support our assumptions about multimodal recognition of intention of interaction. The body pose, especially the shoulders orientation were shown to be relevant for intentionality detection. Spatial information coming from the audio and telemeter streams are also important. The position and size of detected faces in the video confirms that facing the robot is a sign of intentionality of engagement.

These evaluations were conducted using an a priori selection of 32 features. This selection was inspired by our literature searches in human-robot interaction, social sciences and cognitive science fields. Results are improved, but can we conclude that our results are generic enough? For instance, we replaced all skeleton information by Schegloff metrics. Even if results validate our hypothesis, we need to check if we do not have interesting information in the left aside 67 other features.

From the results[5], we see that LEAVE_INTERACT is never well classified in the telemeter condition. In the multimodal condition, the precision and recall scores get slightly improved. Anyway, LEAVE_INTERACT is most of time confused with SOMEONE_AROUND. Several explanations may enlighten this result. When someone is interacting with the Kompaï, he is close to the robot. We have spatial features computed from laser telemeter but no information about his body from the Kinect (see table 7). A more intuitive point can be that less social signals are expressed when leaving interaction. For closeness reasons, the INTERACT class presents also low classification results. We do not actually need to detect it: intention of engagement is a prior state to interaction.

In preliminary conclusion, we can say that our primary hypothesis is validated: multimodality can improve engagement detection on a companion robot with embedded sensors. Nevertheless, other experiments must be conducted with a three classes approach (NO_ONE, SOMEONE_AROUND and WILL_INTERACT) and all the available features.

5.3 Second experiment

In this second experiment, we will tackle our classification task in regards to the lessons learned in 5.2.4.

5.3.1 New 3 labels corpus

Validation of our 3 classes We need to validate our hypothesis about the confusion of the LEAVE_INTERACT and SOMEONE_AROUND classes. We conducted clustering experiments using the k-means algorithm. We wanted to check if it is difficult to separate these 2 classes in the features space. K-means is an algorithm that produces the best clustering knowing the number of wanted clusters. We ran K-means from 2 to 1500 clusters with our 7-features set and checked the distribution of each feature vectors in these clusters. In all clusterings done, there is no significant separation between our 2 classes, i.e. the feature vectors of each class are equitably distributed among clusters. We did clustering with every feature set up to our 32 features and distribution remained diffused. This result also corroborates that it is likely to say that either people do not express strong social signals when they leave interaction with the robot or that in our hardware setup, we cannot compute them.

New labeling We modified our labeling using this time only our 3 classes (see discussion 5.2.4). All frames from the INTERACT class were removed and all instances of LEAVE_INTERACT were replaced by SOMEONE_AROUND. From the remaining 124282 frames, using k-cross folding (see 5.1) with $k = 10$, we

[5] As we did a *stratified* k-fold-cross validation, we have many confusion matrices. Presenting one will not correspond to the k-fold-cross validation result (table 4 and 5), showing all is not possible.

Class	Precision	Recall
WILL_INTERACT	0,95	0,93
NO_ONE	0,93	1,00
SOMEONE_AROUND	0.99	0,84

Table 6: Results of multimodal SVM 5-class classification using Sklearn with our 3 classes.

computed train and test sets respectively of 111854 and 12428 frames. Repartition of each class is given in the figure 13.

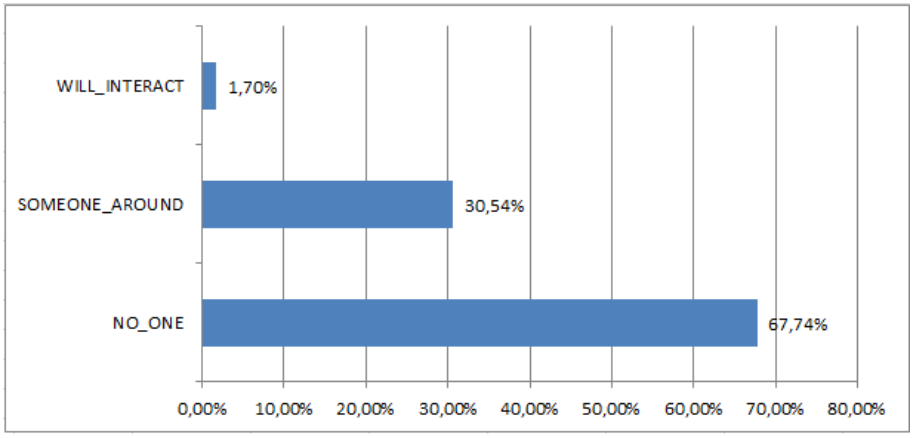

Figure 13: Percentage of frames per event with 3-classes labeling

5.3.2 Classification

Using the new dataset, we re-performed the classification process using SVM. Results are shown in the table 6. We can see that we have an overall improvement in our results on the 3 remaining classes, mainly on the recall score. We increased both scores on the SOMEONE_AROUND class.

Training on this data was more successful, but new experimentation in other conditions, in other places, with different lighting conditions, with more participants has to be done in order to conclude definitely on these results.

5.3.3 Feature selection among all available features

We re-ran our experiment using the MRMR technique to select in our total 99 features set the most relevant ones. Results in this case differ from the first experiment. Some intuitive features, like the *facing coefficient* and the *skeleton distance*, were selected. Surprising features appear among the more relevant ones. Indeed, the most important feature is the *right ankle x* position. This fact is peculiar when one knows that more than 70% of the frames have no information about a skeleton. Moreover, many skeletons have noisy feet information due to the position of the robot in the living lab. During our recordings, participants passed very closed to the robot, as we wanted them to do (see SOMEONE_AROUND example in A.1). In this case, the mounted Kinect did not manage to compute confident 3D feet position.

The MRMR technique is not efficient on this task. For a 99 features space, we may not have enough data to determine reliable metrics (mutual information, correlation, t-test/F-test...) used by the algorithm. Nevertheless, this experiment confirms that selecting human readable features inspired from social and cognitive sciences could be an alternative method for feature space reduction.

6 Conclusion

Psychologists working on the acceptance of a robot by the elderly and people with disabilities have pointed out the need for more natural and acceptable interactions with the companion robot in the home environment. A starting engagement is the first step of interaction. It corresponds to the phase preceding the actual interaction, when the user implicitly signifies his wish to interact. For a companion robot, the skill to detect engagement , and thus, to anticipate human will is a key feature to make it socially acceptable.

Our goal was to evaluate and measure humans' cues for engagement in an interaction with a robot. They were classically detected using the position and the speed of the user. We have shown the limits, in term of recall, in performance of this technique, confronting it with realistic scenarios of engagement towards a robot in a home environment. Indeed, the proximity of the user with the companion robot is not a sufficient criteria when predicting the engagement. Spatial based detection of intention of interaction used in previous approaches gave good results in lab environment. However, the congestion of home-environment leads to situations where humans pass close to the robot without the will of interaction and where spatial based detection gives false positive responses.

Having built realistic scenarios involving the interaction of participants with the Kompaï robot, we have collected sensory data of various engagement sequences. Several features were computed over multiple modalities. From the video, we detected the size and position of the face in the image. The skeleton data gave us clues to compute body poses. The audio was used for the sound localization and for the speech activity recognition. Telemeters gave us an estimation of the position and the speed of the pedestrians.

A cross-fold validation allowed us to segment our dataset into training and testing sets. These subsets where used by two different classifiers, a Neural Networks and a Support Vector Machine. These classifiers, trained on multimodal and telemeters features set, gave better performances for the multimodal condition. This fact showed that spatial and speed features used in related works are not enough in a home environment. Multimodality improved the recall of the engagement detection significantly, which was the hypothesis of this research.

6.1 Key points

Transposing social and cognitive sciences results and using human readable features can be an alternative approach for feature selection. Using this methodology, we enhanced spatial information with the selected body related and acoustic features and get better detection scores, notably in terms of recall. As far as we know, we validated experimentally for the first time, that shoulder pose rotation, a metric from Schegloff's research in Sociology, is of importance to the detection of intention of engagement.

The high correlation between the features also made the classification more difficult. On one hand, the Minimum Redundancy Maximum Relevance (MRMR) feature selection algorithm helped us get to a set of measurable multimodal features sufficient to detect intention of interaction towards a robot based on human selected features set. On the other hand, trying to deals with all 99 available features to elicit the more relevant ones fails. New experiments need to be conducted with more variability in order to improve the scores of the intention detector using the reduced feature space.

Current work about high-level features fusion and analysis is ongoing. We want to remove some artifacts that penalize feature selection and classification algorithms. For instance, we are combining face detection, skeleton tracking and depth data to improve feature association. These features now belong to the same user, i.e. when we compute Schegloff and face features they refer to the tracked pedestrian. As far as we can say from our preliminary experimentation, doing so, results are improved but not significantly for the multi-user scenario.

6.2 Impact of this research

With this research we provide deeper knowledge about meaningful features that can facilitate robot's social abilities. We computed new features inspired from the literature in social sciences, notably the Schegloff's features. A ranked list of 32 most relevant features for our starting engagement can be found

in A.2. This presented work gives design guidelines and praise multi-modal sensor embedding on robot to facilitate human-robot interaction.

The feature selection process depicted in this article was inspired from research on genome [26]. Principal Component Analysis (PCA) and the Linear Discriminant Analysis (LDA) did not provide significant classification improvement. The MRMR algorithm however showed that selecting features can be more relevant than combining them. The direct ranking provides information about meaningful features for a classification task, for example. The feature selection method can be applied in many contexts where space reduction is of interest.

This work is one more step into the use of multimodality for social signal processing applied to human-robot interaction. Multimodality can be very useful in decoding and recognising affect signals and hence in improving the human-robot relationship. With more and more powerful embedded systems deployed on robots, we can expect such multimodal detection to be generalized in real-time and to allow robots to predict intentions of the users. The prediction of the engagement is a first step towards a smoother and socially acceptable human-robot interaction.

7 Acknowledgments

The author would like to thank the Robosoft company for loaning us the Kompaï robot, Inria and French Ministry of Education and Research for their support.

8 References

References

[1] Michael Argyle. *Bodily Communication*. Methuen & Co Ltd, 1975.

[2] V. Aubergé, Y. Sasa, T. Robert, N. Bonnefond, and B. Meillon. Emoz: a wizard of oz for emerging the socio-affective glue with a non humanoid companion robot. *Workshop on Affective Social Speech Signal (WASSS2013), sattelite event of Interspeech 2013*, aug 2013.

[3] Wafa Benkaouar and Dominique Vaufreydaz. Multi-Sensors Engagement Detection with a Robot Companion in a Home Environment. pages 45–52, Vilamoura, Algarve, Portugal, October 2012.

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

[5] Jennifer L Burke, Robin R Murphy, Erika Rogers, Vladimir J Lumelsky, and Jean Scholtz. Final report for the darpa/nsf interdisciplinary study on human-robot interaction. *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on*, 34(2):103–112, 2004.

[6] Ginevra Castellano, André Pereira, Iolanda Leite, Ana Paiva, and Peter W McOwan. Detecting user engagement with a robot companion using task and social interaction-based features. pages 119–126. ACM, 2009.

[7] Tsuhan Chen and Ram R Rao. Audio-visual integration in multimodal communication. *Proceedings of the IEEE*, 86(5):837–852, 1998.

[8] Juan Fasola and Maja J. Matari. Socially assistive robot exercise coach: Motivating older adults to engage in physical exercise. In Jaydev P. Desai, Gregory Dudek, Oussama Khatib, and Vijay Kumar, editors, *Experimental Robotics*, volume 88 of *Springer Tracts in Advanced Robotics*, pages 463–479. Springer International Publishing, 2013.

[9] David Feil-Seifer and Maja J Mataric. Defining socially assistive robotics. In *Rehabilitation Robotics, 2005. ICORR 2005. 9th International Conference on*, pages 465–468. IEEE, 2005.

[10] David Fischinger, Peter Einramhof, Walter Wohlkinger, K Papoutsakis, Peter Mayer, Paul Panek, T Koertner, S Hofmann, Antonis Argyros, Markus Vincze, et al. Hobbit-the mutual care robot. In *Workshop-Proc. of ASROB*, 2013.

[11] James Glasnapp and Oliver Brdiczka. A Human-Centered Model for Detecting Technology Engagement. *Human-Computer Interaction*, LNCS 5612:621–630, 2009.

[12] Edward Twitchell Hall and Edward T Hall. *The hidden dimension*, volume 1990. Anchor Books New York, 1969.

[13] Peng Hanchuan, Long Fuhui, and Ding Chris. Feature Selection Based on Mutual Information : Criteria of Max-Dependency, Max-Relevance and Min-Redundancy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(8):1226–1238, 2005.

[14] Patrick Holthaus, Karola Pitsch, and Sven Wachsmuth. How Can I Help? *International Journal of Social Robotics*, 3(4):383–393, September 2011.

[15] H. Huettenrauch, K. Severinson Eklundh, A. Green, and E.A. Topp. Investigating spatial relationships in human-robot interaction. pages 5052–5059, 2006.

[16] Steven H. Kaminski. *Communication Models*. 2002.

[17] Kostas Karpouzis, Amaryllis Raouzaiou, Athanasios Drosopoulos, Spiros Ioannou, Themis Balomenos, Nicolas Tsapatsoulis, and Stefanos Kollias. Facial expression and gesture analysis for emotionally-rich man-machine interaction. *3D modeling and animation: synthesis and analysis techniques*, pages 175–200, 2004.

[18] Heather Knight. Eight lessons learned about non-verbal interactions through robot theater. *Social Robotics*, pages 42–51, 2011.

[19] Seongyong Koo, Dong-soo Kwon, and Extended Kalman Filtering. Recognizing Human Intentional Actions from the Relative Movements between Human and Robot. *Nonlinear Dynamics*, pages 939–944, 2009.

[20] Nicole C Krämer, Sabrina Eimler, Astrid von der Pütten, and Sabine Payr. Theory of companions: what can theoretical models contribute to applications and understanding of human-robot interaction? *Applied Artificial Intelligence*, 25(6):474–502, 2011.

[21] Peter Krauthausen and Uwe D Hanebeck. Situation-specific intention recognition for human-robot cooperation. pages 418–425. Springer, 2010.

[22] Marwa Mahmoud, Tadas Baltrušaitis, Peter Robinson, and Laurel D Riek. 3d corpus of spontaneous complex mental states. *Affective computing and intelligent interaction*, pages 205–214, 2011.

[23] Jerome Maisonnasse, Nicolas Gourier, Oliver Brdiczka, Patrick Reignier, and James Crowley. Detecting privacy in attention aware system. IET, IET, jul 2006.

[24] Ross Mead, Amin Atrash, and Maja J Matarić. Proxemic Feature Recognition for Interactive Robots : Automating Metrics from the Social Sciences. pages 52–61, 2011.

[25] Albert Mehrabian. Pleasure-arousal-dominance: A general framework for describing and measuring individual differences in temperament. *Current Psychology*, 14(4):261–292, 1996.

[26] Piyushkumar A Mundra and Jagath C Rajapakse. Svm-rfe with mrmr filter for gene selection. *IEEE Transactions on NanoBioscience*, 9(1):31–37, 2010.

[27] Ryota Ooko, Ryo Ishii, and Yukiko I Nakano. Estimating a user?s conversational engagement based on head pose information. *Intelligent Virtual Agents*, pages 262–268, 2011.

[28] Maja Pantic and Leon JM Rothkrantz. Toward an affect-sensitive multimodal human-computer interaction. *Proceedings of the IEEE*, 91(9):1370–1390, 2003.

[29] Sylvie Pesty and Dominique Duhaut. Artificial companion: building a impacting relation. In *Robotics and Biomimetics (ROBIO), 2011 IEEE International Conference on*, pages 2902–2907. IEEE, 2011.

[30] Rosalind W. Picard. *Affective Computing*. International Series in Experimental Social Psychology. The MIT Press, 2005.

[31] Ming-Zher Poh, Daniel J McDuff, and Rosalind W Picard. Non-contact, automated cardiac pulse measurements using video imaging and blind source separation. *Optics Express*, 18(10):10762–10774, 2010.

[32] James M Rehg, Kevin P Murphy, and Paul W Fieguth. Vision-Based Speaker Detection Using Bayesian Networks. *Pattern Recognition*, (Cvpr 99):110–116, 1999.

[33] Charles Rich, Brett Ponsler, Aaron Holroyd, and Candace L. Sidner. Recognizing engagement in human-robot interaction. *2010 5th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 375–382, March 2010.

[34] Jyotirmay Sanghvi, Ginevra Castellano, Iolanda Leite, André Pereira, Peter W McOwan, and Ana Paiva. Automatic analysis of affective postures and body motion to detect engagement with a game companion. pages 305–311. IEEE, 2011.

[35] Jyotirmay Sanghvi, Ginevra Castellano, Iolanda Leite, André Pereira, Peter W McOwan, and Ana Paiva. Automatic analysis of affective postures and body motion to detect engagement with a game companion. *6th ACM/IEEE International Conference on Human-Robot Interaction (HRI2011)*, pages 305–311, 2011.

[36] Emanuel A Schegloff. Body Torque. *Social Research*, 65(3):535–596, 1998.

[37] C. Schroeter, S. Mueller, M. Volkhardt, E. Einhorn, C. Huijnen, H. van den Heuvel, A van Berlo, A Bley, and H.-M. Gross. Realization and user evaluation of a companion robot for people with mild cognitive impairments. pages 1153–1159, May 2013.

[38] Scikit-learn. `http://scikit-learn.org/stable/`.

[39] Candace L Sidner, Christopher Lee, and Neal Lesh. Engagement rules for human-robot collaborative interactions. *IEEE International Conference On Systems Man And Cybernetics*, 4:3957–3962, 2003.

[40] De Silva and Palmerston North. Audiovisual Recognition. pages 649–654, 2004.

[41] SSPNet. Social signal porcessing network `http://sspnet.eu/2010/04/semaine-corpus/`, 2010.

[42] SSPNet. Social signal porcessing network `http://sspnet.eu/`, 2012.

[43] Karim A. Tahboub. Intelligent Human-Machine Interaction Based on Dynamic Bayesian Networks Probabilistic Intention Recognition. *Journal of Intelligent and Robotic Systems*, 45(1):31–52, March 2006.

[44] Dominique Vaufreydaz, Rémi Emonet, Patrick Reignier, and Reignier Patrick 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*, 2006.

[45] David Vernon, Claes Hofsten, and Luciano Fadiga. *A Roadmap for Cognitive Development in Humanoid Robots*, volume 11 of *Cognitive Systems Monographs*. Springer Berlin Heidelberg, Berlin, Heidelberg, 2011.

[46] Alessandro Vinciarelli, Maja Pantic, and Hervé Bourlard. Social signal processing: Survey of an emerging domain. *Image and Vision Computing*, 27(12):1743–1759, 2009.

[47] Michael L Walters, Kerstin Dautenhahn, René Te Boekhorst, Kheng Lee Koay, Dag Sverre Syrdal, and Chrystopher L Nehaniv. An empirical framework for human-robot proxemics. *Procs of New Frontiers in Human-Robot Interaction (2009)*, 2009.

[48] Lin Wang, Pei-Luen Patrick Rau, Vanessa Evers, Benjamin Krisper Robinson, and Pamela Hinds. When in Rome: the role of culture & context in adherence to robot recommendations. *ACM*, pages 359–366, March 2010.

[49] Weka3. Data mining software and toolkit in java `http://www.cs.waikato.ac.nz/ml/weka/`.

[50] WillowGarage. Open cv library `http://opencv.org/`.

[51] C. Wolf, J. Mille, E. Lombardi, O. Celiktutan, M. Jiu, M. Baccouche, E Dellandrea, C.-E. Bichot, C. Garcia, and B. Sankur. The liris human activities dataset and the icpr 2012 human activities recognition and localization competition. March 2012.

[52] Zhihong Zeng, Maja Pantic, and Thomas S Huang. Emotion recognition based on multimodal information. *Affective Information Processing*, pages 241–265, 2009.

[53] Huijing Zhao and R. Shibasaki. A novel system for tracking pedestrians using multiple single-row laser-range scanners. *IEEE Transactions on Systems, Man and Cybernetics, Part A: Systems and Humans*, 35(2):283–291, 2005.

APPENDIX

A.1 Samples of the corpus

The table 7 shows samples of the dataset that we propose for 4 of the events WILL_INTERACT, INTERACT, LEAVE_INTERACT and SOMEONE_AROUND. As one can see the event LEAVE_INTERACT can be quite confusing with the INTERACT state.

WILL_INTERACT

INTERACT

LEAVE_INTERACT

SOMEONE_AROUND

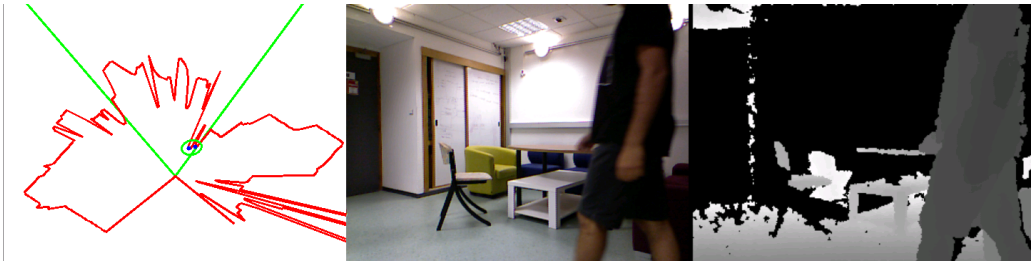

Table 7: Samples of data recorded with the Kompaï equipped with a Kinect sensor and a laser telemeter. For each view, one can find spatial information at left, rgb camera view with face detection in the middle and depth camera with people and skeleton detection.

A.2 Ordered list of 32 most relevant features

Using the Minimum Redundancy Maximum Relevance (MRMR) algorithm, we ranked 32 features (see section 5.2.3). In the following table, the reader must note that *cible_* prefix identifies pedestrian related features. Other body features are related to Schegloff work (see section 4.3.1). Features not listed in table 8 are depicted in section 4.

Order	**Short name**	**Unit**	**Description**
1	shoulderPose_rot	*radian*	Rotation of the shoulder
2	cible_dx	$meter.seconde^{-1}$	Speed in x of pedestrian
3	cible_y	*meter*	position on Y axis of pedestrian
4	face_size	*pixel*	Size of face in the RGB frame
5	face_x	*pixel*	Lateral position of the face
6	beam	*radian*	Activated audio beam
7	angle	*radian*	Audio localization (azimut)
8	hipPose_x	*meter*	Hip X attribute
9	hipPose_y	*meter*	Hip Y attribute
10	hipPose_rot	*radian*	Hip rotation angle
11	face_y	*pixel*	Height of the face
12	sad_event	Speech/Not speech	Speech activity detection tags
13	stancePose_rot	*radian*	Stance rotation
14	torsoPose_rot	*radian*	Torso rotation
15	shoulderTorque	*radian*	Shoulder torque
16	shoulderPose_y	*meter*	Shoulder Y attribute
17	source_confidence	$[0; 1]$	Audio localization confidence
18	torsoTorque	*radian*	Torso torque
19	stancePose_z	*meter*	Stance Z attribute
20	skl_dist	*meter*	Distance of the tracked skeleton
21	cible_x	*meter*	Position on X-axis of pedestrian
22	hipTorque	*radian*	Hip torque
23	torsoPose_y	*meter*	Torso Y attribute
24	torsoPose_x	*meter*	Torso X attribute
25	shoulderPose_x	*meter*	Shoulder X attribute
26	stancePose_x	*meter*	Stance X attribute
27	cible_dy	$meter.seconde^{-1}$	Speed on Y-axis of pedestrian
28	cible_dist	*meter*	Distance of the pedestrian
29	torsoPose_z	*meter*	Torso Z attribute
30	stancePose_y	*meter*	Stance Y attribute
31	hipPose_z	*meter*	Hip Z attribute
32	shoulderPose_z	*meter*	Shoulder Z attribute

Table 8: MRMR algorithm output on the 32 features set.