Multi-Sensors Engagement Detection with a Robot Companion in a Home Environment

Wafa Benkaouar, Dominique Vaufreydaz

Multi-sensors engagement detection with a robot companion in a home environment

Wafa Benkaouar[1] and Vaufreydaz Dominique[1,2]
[1] PRIMA Team - Inria/LIG/CNRS, 655, avenue de l'Europe - 38334 Saint Ismier Cedex
[2] Université Pierre Mendès-France, BP 47 - 38040 Grenoble Cedex 9
Wafa.Benkaouar@inria.fr, Dominique.Vaufreydaz@inria.fr

***Abstract*— Recognition of intentions is an subconscious cognitive process vital to human communication. This skill enables anticipation and increases the quality of interactive exchanges between humans. Within the context of engagement, i.e. intention for interaction, non-verbal signals are used to communicate this intention to the 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. Classically, the human position and speed, the human-robot distance are used to detect the engagement. Our hypothesis is that this method is not enough in the context of a home environment. The chosen approach integrates multimodal features gathered using a robot equipped with a Kinect. The evaluation of this new method of detection on our corpus collected in spontaneous conditions highlights its robustness and validates use of such a technique in real environment. Experimental validation shows that the use of multimodal sensors gives better precision and recall than the detector using only spatial and speed features. We also demonstrate that 7 multimodal features are sufficient to provide a good engagement detection score.**

I. Introduction

Social signal processing and affective computing have emerged as new areas of Computer Sciences over the last ten years (R. Picard [1]). These new areas explore the multimodal aspect of the human communication in order to develop more natural interaction between humans and computers or robots.

Speech is an important channel for communication and requires signal processing as well as semantics and linguistics domain. In addition to the semantics of speech, emotions, and the inner goals of humans are conveyed by other channels: body, gesture, etc. The research community is increasingly interested in this non-verbal (NV) communication. Recognition of intention is a basic skill acquired by infants early in their development. Vernon in [2] 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. Human cognition has a high part of anticipation, allowing to read the intentions, and guessing the goal in order to react quickly to some stimulus.

Companion robots should also be able to detect the intentions of humans in order to adapt their behavior during interactions with humans. For natural human-robot interaction, the intention reading of the behavioral cues from an individual is fundamental.

Our goal for this research is to investigate techniques to detect and recognize signals for non-verbal communication reflecting intentions and in particular the engagement of a human with a robot. We define engagement as the phase during which one expresses, with NV cues, the intention of an interaction. 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 familiar to interacting with robots as shown in [3]. Engagement is fundamental for communication between human users and interactive robots.

Classically, the criterion for a user's engagement are spatial and speed information between the user and the communicant interface [4]. These studies made a simple assumption: 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-like interfaces, but for an assistant living robot in real-life, close distance does not necessary 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, a robot in order to have more human acceptable behavior should be able to detect when it is about to be solicited, and to anticipate this interaction. In the context of a companion robot the proximity of the robot with a person should not be a continuous trigger for engagement. Other criterion can be taken into account such as the posture, the sound and other features described below.

We propose a multimodal approach for detecting engagement using the Kinect© sensors from Microsoft [5] to improve re-usability, and to enable us to build a detector deployable in real-life situations. From literature, in particular 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 multimodal dataset gathered from the robot point of view. Optimization of the acquisition process was needed to limit information loss and to facilitate the synchronization of the multimodal data. This corpus offers a realistic framework to

test our hypothesis.

Evaluation using Multi-class Support Vector Machine and Artificial Neural Networks techniques to classify the features computed from the dataset have given significantly better results in the multimodal condition when compared to a unimodal spatial condition. We show that the spatial and speed features can be improved for engagement detection in a home environment. A subset of 7 multimodal features is proposed for the engagement detection task.

In the following sections, we first develop an overview of the approaches concerning engagement models in cognitive sciences and human-robot interaction. Then, we describe the recording of a robot centered corpus in a home environment and features we can extract from it. Finally, classification and space reduction evaluations are depicted to validate our hypothesis.

II. From Cognitive Sciences to Human-Robot Interaction

Humans are endowed by range of abilities called social intelligence [6]. They include the ability to express and recognize social signals produced during social interactions like agreement, politeness, empathy, friendliness, conflict, etc. They are coupled with the ability to manage these signals in order to get along with others while winning their cooperation.

An intelligent agent is commonly defined as an agent who perceives, learns, and adapts to the world. Social signals are manifested through a multiplicity of non-verbal behavioral cues including facial expressions, body postures and gestures, vocal outbursts like laughter, etc. , which are aimed to be analyzed by signal processing technologies, or automatically generated by synthesis technologies.

Social sensible computer systems and devices which are able to adapt their response to social signals in a polite, non-intrusive, or persuasive manner, in real-time, are likely to be perceived as more natural, efficient and trustworthy. In the context of assistance to personal living in a home environment, social adequacy seems to be crucial for the acceptance of a robot companion.

A. Intentionality in Human-Machine Interaction

Recognition of humans' intentions, goals and actions is important in the improvement of non verbal human-robot cooperation. Intention recognition is defined in [7] by the process of estimating the force driving humans actions based on noisy observations of humans' interaction with his environment. The DARPA/NSF in its final report on Human-Robot Interaction [8] recommends to improve the models of human-robot relationship and in particular to work on the intentionality issue.

In his study, Knight [9] points the importance for a robot to convey and to detect intentionality. It helps to clarify current activity and to anticipate the goals. Learning from the human engagement, the robot would be able to anticipate the interaction and also to learn adequate moments when the robot itself can engage an interaction. In [10] engagement is defined as the process by which two (or more) participants establish, maintain and end their perceived connection during interactions they jointly undertake.

Different modalities are 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 and gesture, the physiologic aspects, clothing, gender, age, etc. We focused on non-invasive aspect of social signal perception and present the modalities used in order to detect engagement.

B. Body Pose and Proxemic features

A way of detecting engagement would be to consider only proxemic metrics. Classical features included in proxemic features are the relative position of the individual to the robot and their relative speed. For a collaboration to be successful, the distance between the robot and the human should be optimum and the speed controlled. In [4], it is proposed to recognize intentional actions using relative movements of a human to a robot. Koo uses an Infrared sensor embedded on the robot to track and estimate the velocity of a person. He then infers intentional actions such as approach and depart using Hidden Markov Models (HMM) and position dependent model.

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 [11]. There is no consensus on the meaning and the emotional characteristics of a posture. Psychologists such as Hall, Mehrabian [12] and Schegloff [13] have proposed some metrics that have been used in computer assisted analysis of posture.

Posture is difficult to measure and evaluate using computer vision. Nevertheless, with the apparition of the Kinect sensor and other real-time 3D pose reconstruction techniques, we are able now to evaluate the pose of a person.

C. Audio Features

Pantic in [14] lists some features into 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. Considering the recognition of the engagement in an interaction, only few papers in the literature use audio features in a multimodal frame. [15] proposes 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 able to localize roughly a sound source. Sound spatialization is not often used for affect detection, but [16] invokes its interest in attention or focus estimation.

D. Facial Features

Concerning the engagement, the orientation of the head and the gaze seem to be crucial. As shown in [17] 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 seems to be a sign of attention.

III. A Corpus for Engagement with a Robot

A part of this study was to record a multimodal dataset including interaction with a robot companion enhanced with a Kinect device.

This section presents the method used to build the dataset needed to test our hypothesis.

A. The need for a dataset

In the context of a companion robot, we want to work with consumer devices in a natural environment. Even though the tendency is to use more and more physiological sensors (such as R. Picard's pulse bracelet Cardiocam, etc.), physiological devices are still invasive and expensive for the users to be released widely.

In order to evaluate our hypothesis, we confronted it to data. In the context of robot companion, the sensors considered are commonly microphones, video sensors, depth sensors, lasers... There exist datasets in the field of social signals processing dealing with non-verbal communication using multi sensors. Available datasets for affect recognition are unfortunately more often for face-to-face interaction with persons sitting and interaction with the speech only. The SSPNet association provided the SEMAINE-DB dataset [18] where several persons have been recorded in a face-to-face speech interaction. This database is suitable for a desktop environment for interaction with virtual communicant agent. Unfortunately, this dataset suits less human-robot interaction, 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 characteristics. Other corpora exist that use the Kinect sensors and 3D information, such as [19] which presents a Cam3D dataset centered on facial and hand movement associated with audio recording. Yet, the proposition of a robot centered dataset for multimodal social signal processing has not been made.

1) Kompai robot: The Kompai robot has been lended by our partner Robosoft[1], allowed us to record our corpus. The Kompai robot, Figure 1, aims at helping elders and dependent persons. It is composed of a RobuLAB mobile platform containing the wheel actuators, obstacle detection system, manual remote control facilities, etc. The mobile platform is topped by a tablet serving as interface with the user, a pair of microphones, a motorized webcam and a speaker, to which we added a Kinect sensor.

In our recordings, we gathered every sensor available like the head-mounted webcam of the robot used to record videos during the experiment.

2) Kinect Sensor: The Kinect sensor is composed of several components which are represented in the Figure 2. Advantages of using such a sensor are its consumer price and its growing utilization in computer vision assisted system. During corpus gathering, we recorded several streams from the Kinect:

- *Depth Camera (using Infrared laser)*: the depth range is limited from 80 centimeters to 4 meters with a 2 millimeters accuracy.
- *Skeleton Tracking*: the Kinect supports up to two skeletons being tracked at the same time. Only the tracked skeletons with a high confidence score are stored in the dataset.
- *RGB Camera*: the resolution of the RGB image is 640x480 pixels by default. The RGB horizontal field of view is of 62.0 degrees.
- *Microphone Array*: the array is composed of four aligned microphones. It provides an angle of a detected sound with a confidence in the Kinect reference frame. It also outputs the more stimulated beam by the sound source.

Fig. 1. The Kompai Robot from Robosoft.

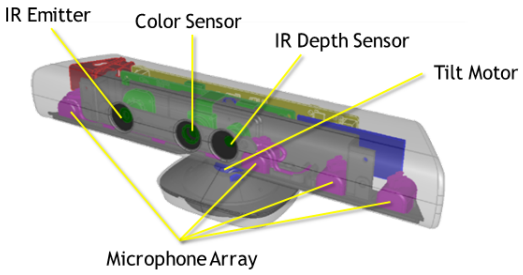

Fig. 2. Components of the Kinect Sensor [5]

B. Features extraction

The recorded data are presented in the table I. Some of them were analyzed to extract features for the engagement detection.

[1] http://www.robosoft.fr/

Data	Sensor	Maximal Frame Rate
Telemeters distances	Kompai	12.5Hz
Ultraound distances	Kompai	12.5Hz
Audio	Kinect	16kHz
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 I

DATA RECORDED, ASSOCIATED SENSORS AND FRAME RATE

1) Features selection: Using all available sensors (see previous section), we must define which features to extract from the data. Looking at the literature, we decided to compute the following features:

- Using the laser telemeter, we can extract, in the frame of the robot, the *x* and *y* position, the *dx* and *dy* speed, and *dist*, the distance to the robot. These features are computed using a background subtraction on the telemeter input and a Kalman Filter to track moving people. This set of features will be, as expressed formerly in the article, our comparison point with the state-of-the-art technique for engagement detection. We named it the *telemeter* condition.
- Using the microphone array, we can add acoustic features: angle, activated beam and confidence of the acoustic source localization and speech activity detection using [20].
- Considering that facial information are important, we computed using OpenCV [21] in the RGB video stream from the Kinect *face_x*, *face_y* and *face_size* respectively the position and size of the biggest detected face in the image.
- Stance, hips, torso and shoulders positions and relative rotations depicted by [13] are computed from the tracked skeletons[2] and give 19 features.

Finally, our set consists in 32 features from different modalities captured from the Kompai enhanced with a Kinect.

2) Features fusion and synchronization: There are 3 main fusion techniques for multimodal corpora: data fusion, features fusion or decision fusion. Data fusion is more suitable when data are of the same kind (multiple video streams for instance). Second, fusion at the feature level aims to aggregate features extracted from the various sensors together before attempting to classify. Last, using late decision fusion has some advantages. The computational cost of the training is reduced and the strict synchrony of the inputs is not required since they bring complementary information. Its drawback is the expertise needed or the relative empiricism of the final decision fusion. According to [22], features fusion is considered more appropriate for closely temporally synchronized input modalities, such as speech and lip movements. As we considered that all our modalities synchronously express our engagement, we decided to use this method.

The common dimension of all the modalities is the time. As seen in table I, frame rate of inputs are different. We decided to synchronize all features on a fixed frame rate. Data from the Kinect present a variable frame rate when recording all streams and tracking people and skeletons at the same time. Only telemeters information is cadenced at fixed frame rate 12.5Hz using a micro-controller. We synchronized everything using the current value of features at the telemeter events timestamps.

C. Realistic Dataset

R. Picard in [1] gives 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 elicit (activated or acted). Another influence can come from the environment of the recording, and the question here is that are 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 our 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 the measurement 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. For many of the participants, this room is new and this can create some fluctuations in the behaviors.

D. Scenarios

In order to test our hypothesis that the position and speed of the person is not enough to detect engagement, we propose to confront with scenarios where users pass close to the robot but with no intention of interaction. The robot is immobile in a waiting attitude.

We want to detect the pre-interaction phase where participant show social signals of their engagement. We made the assumption that these cues were detectable with the sensor that equipped our version of the Kompai robot.

The data were recorded with two different scenarios performed several times by different participants in a homelike environment with a Kompai. The room is similar to a small flat (Figure 3). It is randomly asked to the participant to enter the room by different doors, perform some realistic actions

[2] As we used the Windows version of the Kinect driver, we did not have specific initialization process for skeleton tracking while recording walking people.

and go out. One of the actions is to interact with the robot. The interaction consists in a small flash game on the tablet PC. The other actions were walking, sitting, or pouring water from the sink. Participants were not aware of our intention to measure their engagement with the robot.

1) Scenario "Passing By": In this first scenario, one participant is ask to go through the room twice by different doors (A), (B) or (C). Figure 3 shows the setting of this scenario.

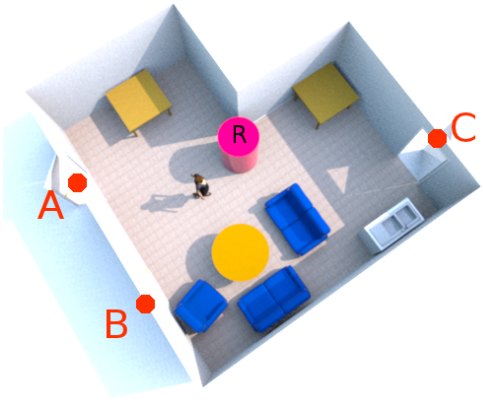

Fig. 3. Scenario 1 *"Passing by"*. A, B and C are access doors. R is the robot.

2) Scenario "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

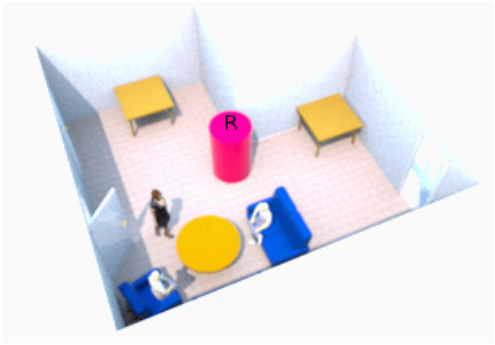

Fig. 4. Scenario 2 *"Playing cards together"*

is used to ask one of the participants to execute an action (gaming interaction with the robot, or using the sink for example). Figure 4 shows this scenario when one of the participant is entering in the room while the other two are already sitting.

E. The dataset in numbers

The recording of the corpus has been made during three sessions of one to two hours. The corpus includes 29 interactions with the robot, made by 15 different participants among more than 50 actions. In real life, all individuals do not express these signals the same way. Some variability has been introduced in the pool of participants. They 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. To randomize the attributions of actions for the participants is also a way of controlling certain pattern in the parasite variables that can appear when experimenting with real data. The duration of the interaction also varies from 2 to 10 minutes according to the participant will.

The total size of the uncompressed data set is around 300 GB with more than 150.000 frames of 32 extracted features.

F. Corpus availability

The corpus is not, for now, available. An enlarged version of the corpus will be recorded with more participants and new scenarios. We plan to release it for the research community.

IV. Automatic labeling

A. Steps of the interaction process

The process of interaction has been described by Sidner and Lee in [10]. They proposed a model in three steps: initiation of interaction (WILL_INTERACT in our labeling), maintenance of interaction (INTERACT) and disengagement (LEAVE_INTERACT). We added two more classes NO-ONE when nobody present and SOMEONE_AROUND when someone is around the robot and does not want to interact.

B. Labeling rules

Our scenarios were defined for helping us in the automatic labeling of the dataset. Before interacting, people are located in blind areas for the telemeters: outside the room or in the game area. Using laser telemeter information, we can detect when someone is moving towards the robot.

The interaction (INTERACT class) appears between the beginning and the end of user clicks on the tablet. The WILL_INTERACT phase preceding the beginning of interaction (first click) is labeled since appearance of a moving object just before the interaction on the robot tablet. In both scenarios, it can be done as people were coming from a blind area for the telemeter: outside for the first scenario, the playing cards area for the second one. LEAVE_INTERACT has been tagged during 5 seconds after the end of interaction. The idea behind this empirical choice is that leaving interaction with the robot is after a short leaving sequence, just like walking away from it. The SOMEONE_AROUND event is labeled when someone is 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.

Automatic labeling has been confronted and validated against manual pre-annotation of recorded sessions.

V. Evaluation

We focus on the engagement detection, i.e. on the WILL_INTERACT class. Other classification results are presented but will not be discussed in this paper.

A. Classification Results

In order to classify our features, we chose to use two kinds of classical classifications: Artificial Neural Networks (ANN) and Support Vector Machines (SVM) techniques. For these two techniques we built and tested two classifiers one for the multimodal dataset (including the whole 32 features) and one for telemeters condition (5 features).

1) Artificial Neural Networks: The Artificial Neural Network is a multi-layered model with perceptrons. We used the Weka [23] toolbox to perform this classification. The use of ANN is common to infer models from observation. In our case, we suppose that our features 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. Results of the ANN classification are presented in the Table III for the telemeters and the Table II for the multimodal feature set.

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 II

RESULTS OF MULTIMODAL NEURAL-NETWORK 5-CLASS CLASSIFICATION.

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 III

RESULTS OF TELEMETER NEURAL-NETWORK 5-CLASS CLASSIFICATION.

First, these results show that the overall precision and recall of the classifier for our classes is slightly better in the multimodal approach. Concerning the engagement class, WILL_INTERACT, the system returns more relevant event as an engagement in the case of the multimodality 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 terms of recall and a low false-positive rate. The Neural Network classifier gave better recall rate in multimodal condition.

2) Multi-Class Support Vector Machine: Tests using Support Vector Machine were done using the Sklearn toolkit [24]. The results of the 5-classes classification using for the multimodal features are presented in Table IV. For the telemeters classification the results are presented by the Table V. We observe, comparing these tables, that the precision and recall scores for the WILL_INTERACT class are significantly improved by the multimodality. Also, for this same class, the False-Positive rate is higher in the case of the telemeters only. In particular, the aim of this detection was to decrease this rate of misclassifying an event as WILL_INTERACT, hence the system has less chance 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 IV

RESULTS OF MULTIMODAL SVM 5-CLASS CLASSIFICATION.

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 V

RESULTS OF TELEMETER SVM 5-CLASS CLASSIFICATION.

B. Feature space reduction

The Minimum Redundancy Maximum Relevance (MRMR) method [25] has been performed in order to highlight the best features for our detection system. Contrary to Principal Component Analysis (PCA) or Linear Discriminant Analysis (LDA), this dimensionality reduction technique has the advantage of selecting the most relevant features instead of building new features by combining the observed ones. MRMR uses mutual information to select features which jointly have the maximal statistical dependency while best characterize the statistical property of a target classification variable. Hence, it could allow discarding some less relevant features in order to optimize the detection of engagement process.

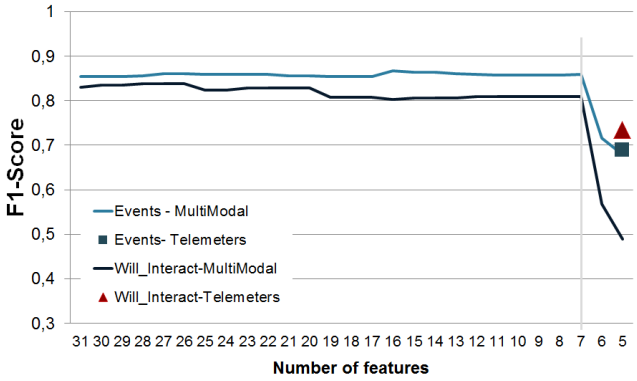

Fig. 5. F1-score evolution while decreasing the number of multimodal features in comparison with the telemeters for all the events and for the WILL_INTERACT event.

Figure 5 shows the impact on the f1-score[3] of the space reduction from 31 to 5 selected features with MRMR.

[3] F1-score is a combination of the *precision* and *recall* values (see `http://en.wikipedia.org/wiki/Precision_and_recall`).

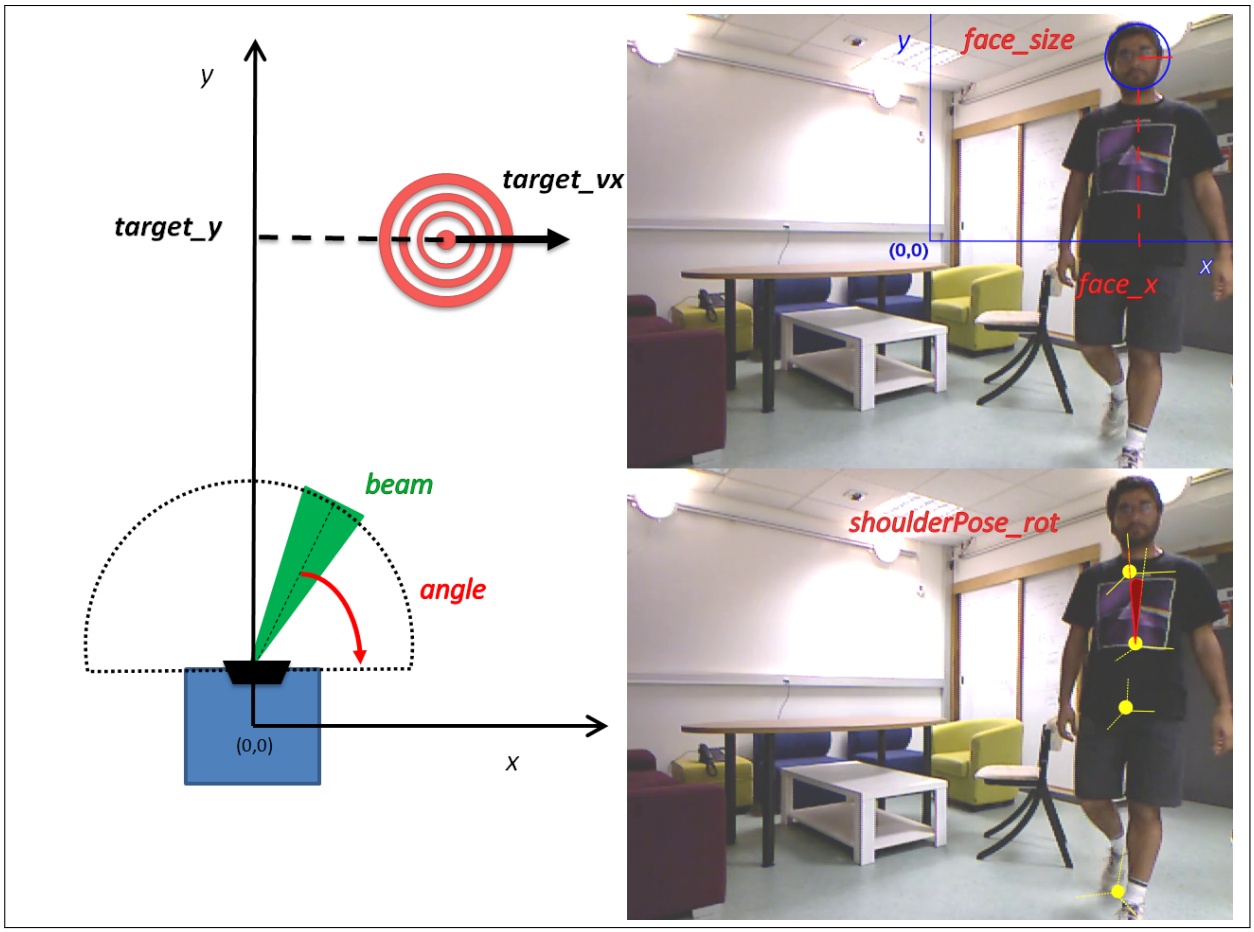

Fig. 6. Minimal multi-modal set with 7 features. The blue square represents the Kompai robot, the black trapezoid the Kinect. *Target_vx* and *target_y* are computed using telemeter information in the robot reference frame. Using the Kinect audio stream, video stream and skeleton tracking, we can respectively extract *angle* and *beam*, *face_size* and *face_x*, and the shoulder rotation (*shoulderPose_rot*).

The performance drops when six features are reached. Before, it remains pretty stable and even non significantly slightly increases along the feature reduction. These results confirm the fact that there are many correlations in the complete feature space. Some of these features seem to be fundamental for a better detection and to keep a precision higher than the telemeters' one.

The first remark on these results is that the 7 highest rated features are coming from heterogeneous modalities. The *shoulderPose_rot* corresponds to the relative orientation of the shoulder in the body, and is extracted from the skeleton information. MRMR classes it as the principal feature. Next, some telemeters information are considered as relevant: *target_vx* and position *target_y*. 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. These features are illustrated in Figure 6.

From these results, our intuitions based on cognitive sciences studies of the engagement recognition are comforted. Indeed, the importance of the body pose, such as the orientation of the shoulder is exposed. Position and size of the face in the image show that the person is facing the robot which a priori confirms its engagement. Some moving criteria complete this features list but not all of them. Distance to the robot, y and dy in the reference frame of the robot are not selected whereas x position and speed are significant in our experiment.

VI. Conclusion and Future Work

In this article, we presented our multimodal approach for engagement detection in a homelike environment with a robot companion enhanced with a Kinect. We recorded a multimodal robot-centered corpus for engagement detection following mono-user and multi-users scenarios. In comparison with the usual spatial features set and using this corpus, we increased precision of multimodal engagement detection respectively from 71% up to 87% with recall staying at 90%.

With feature space reduction technique, we highlight the 7 most relevant multimodal features for engagement detection from our features set. Shoulder rotation, face position and size, user distance and lateral speed, sound localization information were found to be coherent with the results on engagement described in cognitive sciences researches.

With a more powerful embedded system and a computation limited to 7 features, we are currently working on a real-time detection on the robot. Prediction of engagement is a first step toward a smoother human-robot interaction.

REFERENCES

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

[2] 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.

[3] 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.

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

[5] Microsoft Reasearch, "Kinect for windows programming guide,", `http://msdn.microsoft.com/en-us/library/hh855348.aspx`, 2012.

[6] SSPNet, "Social signal processing network,", `http://sspnet.eu/about/`, 2012.

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

[8] J. L. Burke, R. R. Murphy, E. Rogers, V. J. Lumelsky, and J. Scholtz, "Final Report for the DARPA / NSF Interdisciplinary Study on Human Robot Interaction," *IEEE, Transactions on Systems, Man, and Cybernetics - PART C: Applications and Reviews*, vol. 34, no. 2, pp. 103–112, 2004.

[9] 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.

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

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

[12] 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.

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

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

[15] R. Ooko, R. Ishii, and Y. I. Nakano, "Estimating a User's Conversational Engagement Based on Head Pose Information," pp. 262–268.

[16] J. Maisonnasse, *Estimation des Relations Attentionelles dans un Environnement Intelligent*. PhD thesis, Universite Joseph Fourier de Grenoble, 2007.

[17] 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.

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

[19] M. Mahmoud, T. Baltru, P. Robinson, and L. Riek, "3D corpus of spontaneous complex mental states," *Corpus*, 2011.

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

[21] WillowGarage, "Facedetection.", `http://www.opencv.willowgarage.com/wiki.FaceDetection`.

[22] A. Jaimes and N. Sebe, "Multimodal human-computer interaction: A survey," *Computer Vision and Image Understanding*, vol. 108, pp. 116–134, Oct. 2007.

[23] Weka, "Weka 3: Data mining software and toolkit in java.", `http://www.cs.waikato.ac.nz/ml/weka/`.

[24] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[25] 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.