Perceiving user's intention-for-interaction: A probabilistic multimodal data fusion scheme

Christophe Mollaret, Alhayat Ali Mekonnen, Isabelle Ferrané, Julien Pinquier, Frédéric Lerasle

PERCEIVING USER'S INTENTION-FOR-INTERACTION: A PROBABILISTIC MULTIMODAL DATA FUSION SCHEME

C. Mollaret[1,2,3]*, A. A. Mekonnen*[1,2]*, I. Ferrané*[1]*, J. Pinquier*[1]*, F. Lerasle*[2,3]

[1]IRIT, Univ de Toulouse, 118 route de Narbonne, 1062 Toulouse Cedex, France
[2]CNRS, LAAS, 7 avenue du Colonel Roche, F-31400 Toulouse, France
[3]Univ de Toulouse, UPS, LAAS, F-31400 Toulouse, France
{cmollare, aamekonn, lerasle}@laas.fr;{ferrane, pinquier}@irit.fr

ABSTRACT

Understanding people's intention, be it action or thought, plays a fundamental role in establishing coherent communication amongst people, especially in non-proactive robotics, where the robot has to understand explicitly when to start an interaction in a natural way. In this work, a novel approach is presented to detect people's *intention-for-interaction*. The proposed detector fuses multimodal cues, including estimated head pose, shoulder orientation and vocal activity detection, using a probabilistic discrete state Hidden Markov Model. The multimodal detector achieves up to 80% correct detection rates improving purely audio and RGB-D based variants.

***Index Terms*—** Intention Detection, Multimodal Data Fusion, Human-Robot Interaction

1. INTRODUCTION

Developmental psychology and cognitive neuroscience studies suggest humans have an inherent tendency to infer other people's intentions from their actions. This provides an intrinsic ability to understand other people's minds and plays a fundamental role in establishing coherent communication amongst people [1]. Inspired by this, different researchers have been working on detecting user's intention for improved human-machine interaction in general, e.g. [2, 3, 4, 5]. Knowing a user's true intention opens up the possibility to: (1) understand his/her activity at the earliest (before the activity is even complete); (2) constrain the space of possible future actions and provide context [4]; and (3) correctly understand his/her action, for example, in the event of a motor neuron disorder where actions might not reflect true user's intention [6]. Consequently, detecting user's intention has, in recent years, gained significant attention in Human-Robot Interaction (HRI) research. Endowing robots with the ability to understand humans' intentions opens up the possibility to create robots that can successfully interact with people in a social setting as humans. By observing user's intention, a robot can potentially consider implicit commands and user's desires that are not explicitly stated.

In this work, we focus on the specific task of detecting a user's *intention-for-interaction* with a robot. This is very important especially when considering a non-intrusive assistive robot. To paint a picture, consider, for example, an assistive robot that only responds to a user when the user expresses need but otherwise stows away at the corner of a room observing the user without any interference. The robot will approach and interact with the user only when it detects the user's intention, mimicking the actions of a domestic helper. To achieve this, we base our system on studies that report on "how people manifest their intention to interact with another person" – orienting their head and body toward that person and expressing their need vocally [1]. Accordingly, we consider three important cues: user's head orientation, anterior body orientation, and vocal activity. We present a multimodal perceptual system for detecting user's intention-for-interaction with a robot. The proposed system estimates the user's head orientation and shoulder orientation using data acquired from an RGB-D sensor. The latter is considered as an indicator of the user's anterior body orientation. It also determines the user vocal activity using an android device (a smart phone or a tablet) placed casually in the vicinity of the user. Then the outputs from the three systems are fused in a probabilistic Bayesian filter, a Hidden Markov Model (HMM), to provide a posterior estimate on the user's intention-for-interaction. Finally, the user's intention is detected by thresholding the posterior estimate. Although the presented perceptual system is based on a robotic system, it is equally applicable for any generic human-machine interaction system equipped with an RGB-D and audio sensor (as demonstrated in section 5).

This paper is structured as follows: This section continues with related works and highlights our contributions. Section 2 presents the proposed framework in detail. Subsequently, sections 3 and 4 detail the proposed head and shoulder estimation, and vocal activity detection modules respectively. Finally, experiments carried out and obtained results are presented in section 5 followed by conclusions in section 6.

This work was supported by a grant from the French National Research Agency (ANR) under grant number ANR-12-CORD-0003.

Related Works Recently, various works revolving around user intention perception have been burgeoning in the HRI community [6, 4, 7, 8]. The need for understanding people's intention mostly stems from early activity detection [5, 7, 9], context establishment [4, 5], and true intention understanding in case of confusing actions [6]. Intention can be described with several aspects, such as the nature of data (monomodal, multimodal, discret, continuous, etc), the fusion strategy, and finally the applicative context.

Focusing on the inputs for the intention perception detector, several data channels can be distinguished. First, the most obvious should be the head pose and eye gaze estimation as demonstrated in Martinez et al. [6]. A second cue comes straightforward with the context awareness in Clair et al. [4]. Bascetta et al. [9] used an online prediction of user's trajectory, which can be associated with a user's habit. More cues are related with user's body part orientation. Huber [3] based his work on user's feet position and orientation, Kuan et al. [7] used elbow angles and force signals. Only a few paper presented audio features such as [10]. In order to extract all these features, RGB-D cameras and classical cameras are dominantly chosen for tracking, head pose and eye gaze estimation, but sometimes physiological sensors are used such as muscular electomyogram (EMG) and force sensors. Surprisingly, contrary to its pervasive presence, audio sensors/signals have been rarely utilized for intention detection, but rather for user engagement detection in few occasions, e.g., [10]

Evidently, fusing different heterogeneous cues further robustifies the estimation step. When considering multimodal/multi-cue based intention estimation, the considered fusion/inference module plays an important role in robustness. In the literature, the most promising works utilize probabilistic frameworks for fusion and inference. For instance Dynamic Bayesian Networks [11], Hidden Markov Models [9, 5], and generic recursive Bayesian filters [6]. Generally, all intention perception modules relate to safety considerations and improved communication. However, they are used in a large variety of applications, such as: action prediction [4], electric wheelchair's navigation [6], and guiding or resisting a user as part of a rehabilitation process [7]. These kinds of perception modules are even used in smart public display in order to determine the intention to read an advertisement [3]. Based on these insights, the presented multimodal intention-for-interaction detection scheme fuses user head orientation, user anterior body orientation, and audio activity – heterogeneous cues that have not been considered altogether before for detecting intention – in a probabilistic framework. Intermediary outputs from a visual detector, head and shoulder orientations, are filtered using a novel Particle Swarm Optimization based tracker. This step is important as intentionality is based on spatio-temporal patterns, and head pose and shoulder orientation seems to be highly correlated.

Contributions In this paper, we claim to make two core contributions: (i) we propose a probabilistic user's intention-for-interaction estimation framework using Hidden Markov Model (HMM) to fuse visual and audio cues from an assistive robot; and (ii) we integrate the intention detection module in a real robotic platform providing quantitative evaluations of each constituent modules and the complete framework.

2. FRAMEWORK

As highlighted in the introduction, the focus of this work is detecting the intention of a user to interact with an assistive robot – referred as intention-for-interaction detection. This will be considered as the working definition of *intention* through out the rest of this paper.

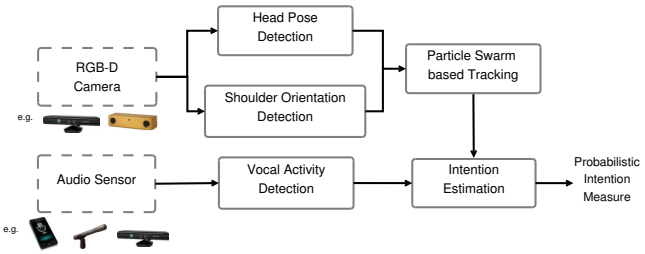


Fig. 1. Proposed framework to estimate user's intention-for-interaction.

Figure 1 shows the proposed framework to estimate user's intention using an RGB-D camera (e.g., kinect, stereo rig) and an audio sensor (e.g., smart phone, tablet, mic, etc). The framework estimates the user's intention based on three important cues: line of sight – inferred from head pose; anterior body direction – determined from shoulder orientation; and speech used to draw attention – captured via vocal activity detection. The head pose detection and shoulder orientation detection modules rely on depth image (detailed in section 3). The detection outputs are further filtered to discard spurious noise and fill-in missing detections using a particle swarm optimization inspired tracker (PSOT). Both, the vocal activity detection (detailed in section 4) and tracker output are considered as observation (measurement) inputs and are fused to provide a probabilistic intention estimate using a Hidden Markov Model (figure 2).

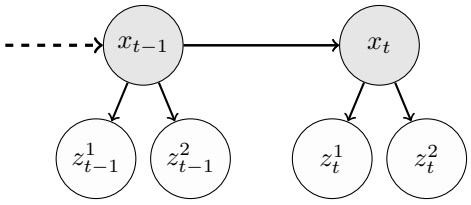


Fig. 2. Probabilistic graphical model used for intention estimation.

The probabilistic graphical model depicted in figure 2 illustrates the relationship between the hidden variables, x_t and x_{t-1}, which are the intention indicators at time t and $t-1$ respectively, and the observation variables z^1_t, z^2_t, z^1_{t-1}, z^2_{t-1}. z^1_t represents the observation from the particle swarm based

tracker that provides estimated head pose (position and orientation) in space, and shoulder orientation with respect to the vertical plane of the camera optical frame in space (yaw). z_t^2, on the other hand, denotes the observation from the vocal activity detection module. Probability distributions associated with each of the observation variables are detailed in sections 3 and 4 respectively.

With the assumption that the observations are conditionally independent given the state (encoded in the graphical model), and making use of Bayes's rule, the posterior probability distribution over the state $P(x_t|Z_{1:t})$ given all measurements upto time t, $Z_{1:t}$, is expressed with equation 1.

$$P(x_t|Z_{1:t}) = \eta P\left(z_t^1|x_t\right) P\left(z_t^2|x_t\right) \sum_{x_{t-1}} P(x_t|x_{t-1}) P(x_{t-1}|Z_{1:t-1}) \quad (1)$$

Where $P(x_t|x_{t-1})$ is the state transition (dynamics) distribution, and η is a normalization factor. Here both x and z^2 are discrete random variables that take on values $\{intent, \neg intent\}$ and $\{vad, \neg vad\}$ respectively, where vad stands for vocal activity detection (see section 4). On the other hand, z^1 takes on continuous values given the tracker output (see section 3).

3. HEAD AND SHOULDER POSE ESTIMATION

This observation is based on: head pose estimation, shoulder orientation estimation, and particle swarm optimization based tracking for filtered estimates.

Head pose estimation This module is based on the works of Fanelli *et al.* [12]. In their work, the authors formulate the pose estimation as a regression problem and use random regression forests on depth images from an RGB-D sensor. This choice is motivated by regression forests capability to handle large training datasets. The regression is based on rectangular features that resemble generalized HAAR-like features. The training is done using the Biwi Kinect Head Pose Dataset [12]. The 6-D state vector, $[\mathrm{x'}_h, \mathrm{y'}_h, \mathrm{z'}_h, \theta'_h, \phi'_h, \psi'_h]$, contains the 3-D head position and the 3 orientation angles relative to the sensor. The claimed precision in the paper is 5.7^o mean error in yaw estimation with 15.2^o standard deviation, and 5.1^o mean error in pitch estimation with 4.9^o standard deviation. Additionally, head pose is detected with $13.4mm$ mean error with $21.1mm$ of standard deviation. This mode works best with close by subjects, subjects placed at a distance of 1.5 to $2.0m$.

Shoulder orientation estimation For this we primarily rely on Openni library [13] which provides a fitted skeleton model of the user based on the depth data. Then, using simple geometry, the user's shoulder orientation is obtained by computing the vector between the left and right shoulder joint pose determined from the fitted skeleton. The shoulder orientation is expressed with respect to the RGB-D sensor parallel optical plane providing a yaw angle θ'_{sh} for the tracking step. Illustrative estimated shoulder and head orientations are displayed in figure 3. Following the Kinect – the specific RGB-D camera used in this work – characteristics and Openni library, the skeleton tracking algorithm works upto a range of 4 meters.

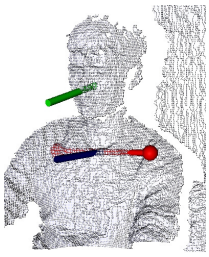


Fig. 3. The head pose is displayed with the green cylinder (head) on the point cloud, while shoulders are displayed in red and their orientation in dark blue (below the neck).

The head and shoulder poses provided by the above two modules are computed frame by frame without any temporal link. It is also possible to have missing estimates from any of the modules at any times. To alleviate that and provide a smoothed continuous estimate, we make use of the following PSOT tracker.

Particle Swarm Optimization based Tracker (PSOT) The PSOT, as detailed here, is a proposed tracker that combines interesting amenities of Particle Swarm Optimization (PSO) [14] with Particle Filter [15], like target dynamic model for improved tracking performance. It is used here to track the head pose and shoulder orientation estimates providing smoothed and improved spatio-temporal estimates.

input : $\hat{s}_{t-1}, \left\{s_{t-1}^{(i)}, p_{t-1}^{(i)}, v_{t-1}^{(i)}\right\}_{i=1}^{N}$

output: $\hat{s}_t, \bar{s}_t, \left\{s_t^{(i)}, p_t^{(i)}, v_t^{(i)}\right\}_{i=1}^{N}$

for $i \leftarrow 0$ **to** N **do**

$r_p, r_g \sim U(0,1)$

$v_t^{(i)} \leftarrow \omega v_{t-1}^{(i)} + \psi_p r_p \left(d\left(p_{t-1}^{(i)}\right) - s_{t-1}^{(i)}\right) + \psi_g r_g \left(d\left(\hat{s}_{t-1}\right) - s_{t-1}^{(i)}\right)$

$s_t^{(i)} \leftarrow s_{t-1}^{(i)} + v_t^{(i)}$

if $f\left(s_t^{(i)}\right) > f\left(p_{t-1}^{(i)}\right)$ **then**

$p_t^{(i)} \leftarrow s_t^{(i)}$

if $f\left(p_t^{(i)}\right) > f\left(\hat{s}_{t-1}\right)$ **then** $\hat{s}_t \leftarrow p_t^{(i)}$

end

end

MAP estimate: $\hat{s}_t$

MMSE estimate: $\bar{s}_t = \sum_{i=0}^{N} \frac{f\left(p_t^{(i)}\right)}{\sum_{j=0}^{N} f\left(p_t^{(j)}\right)} s_t^{(i)}$

Algorithm 1:Particle swarm optimization based tracker.

Contrary to other particle based algorithms, in the presented PSOT, particles interact with each other with a "social" and "cognitive" component in the update step. This behavior leads to a more efficient estimation of target as there is not particle degeneration. Algorithm 1 outlines the utilized PSOT tracker. The particle swarm is indexed with i,

and based on classical PSO terminology, $s_t^{(i)}$, $v_t^{(i)}$, $p_t^{(i)}$ represent the i^{th} particle tracked state in the search space, velocity, and best known tracked state respectively at time t. $f(\cdot)$ is a fitness function (likelihood equivalent in Particle Filters), r_p and r_g are factors that randomly weight social and cognitive terms, whereas ψ_g and ψ_p are constant social and cognitive weights. ω models particle inertia, $\hat{s}_t$ the Maximum A Posteriori (MAP) estimate at time t, and $\bar{s}_t$ the Minimum Mean Square Error (MMSE) estimate.

In this work a random walk dynamic model, i.e. $d(p_t^{(i)}) = p_t^{(i)} + w$ (w is a Gaussian noise), is adopted for $d(\cdot)$ – the filter dynamic model. Given head pose and shoulder orientation in the form of $s_d = [\mathrm{x'}_h, \mathrm{y'}_h, \mathrm{z'}_h, \theta'_h, \phi'_h, \psi'_h, \theta'_{sh}]$ from the head pose and shoulder orientation estimation modules, algorithm 1 is used to determine spatio-temporal posterior point estimates on head pose and shoulder orientation recursively at each time frame t with the state vector of the PSOT tracker represented as $s = [\mathrm{x}_h, \mathrm{y}_h, \mathrm{z}_h, \theta_h, \phi_h, \psi_h, \theta_{sh}]$. A multivariate Gaussian model with a diagonal covariance matrix $diag(\Sigma) = [diag(\Sigma_{position}) \;\; diag(\Sigma_{angles})]$ is used as the observation model in the fitness evaluation.

The distribution $P(z_t^1|x_t)$ is derived based on the tracker output. It is mainly based on $\theta'_h, \phi'_h, \theta'_{sh}$ angles. These angles are represented in such a way that when the user is looking right into the optical frame with their anterior body oriented parallel to the image plane, all angles are 0. With this in mind, $P(z_t^1|x_t)$ is represented as a multivariate normal distribution, i.e., $P(z_t^1|x_t) = \mathcal{N}\left(z_t^1; 0, \Sigma\right)$ with $z_t^1 = [\theta'_h, \phi'_h, \theta'_{sh}]$. The covariance matrix Σ is a diagonal matrix; though not applied here, its values can be varied using the tracked head position.

4. VOCAL ACTIVITY DETECTION

As stated in section 2, audio signal is used for intention detection based on user vocal activity detection. Users have the tendency to talk to a robot when they want its attention. Taking advantage of this, we denote the onset of a vocal activity as one indicator for user intention-for-interaction. In this work we rely on the vocal activity detection module from PocketSphinx C library[1].This algorithm is based on signal energy. It flags the given audio frame as containing speech elements if the signal energy is above a set threshold. Since signal energy is affected by the noise in the environment, the implementation in PocketSphinx does an initial calibration stage so as to best separate signal from stationary noise using a statistical-based noise removal method. Depending on the environment ambient noise (robot noise and room noise), the VAD is estimated properly up to 2 meters. Hence, the user should be within 2 meters from the android based mobile device.

The observation from the VAD module is represented by the random variable z_t^2 at time t taking discrete values $\{vad, \neg vad\}$. Since both z_t^2 and x_t take binary discrete values, the associated likelihood distribution $P(z_t^2|x_t)$ is simply represented by four probability values.

5. EXPERIMENTS AND RESULTS

The presented work is aimed to be deployed on a mobile robot for user's intention-for-interaction detection in HRI setting targeted for the elderly people. The complete framework is implemented using C++ and Python languages as various interacting ROS[2] nodes. It has been integrated on a PR2 mobile robot. The PR2 is a popular robot that has been used as a test bed by many robotic researchers all over the world. It is equipped with various sensors of which we have primarily used the kinect RGB-D sensor mounted on it. Its computing power relies on two Quad-Core i7 Xeon Processors (8 cores) with 24 GB RAM. Audio data is captured using Samsung Note 3 smartphone (android 4.2). The smartphone communicates with PR2 via a common wifi network.

5.1. Evaluation Metrics

For evaluation, we make use of two sets of evaluation metrics: Tracker accuracy metrics, to evaluate the performance of the PSOT tracker, and various metrics to characterize the intention detection performance. For the PSOT tracker, we treat the position estimates and angular estimates separately and compute positional root mean square error (RMSE) in mm and angular RMSE in degrees.

For user's intention detection, we make use of various metrics mosly used in HMM applications in the literature.
True Positive Rate (TPR): The ratio of correct intention detection (in accordance with the ground truth) to that of total intention tagged data frames.
False Alarm Rate (FAR): The ratio of the number of observation data frames, of which the detection output flags an intention where there is none in the ground truth.
Average Early Detection (AED): Given an observation length in time of T, early detection time is the time t the system took to correctly detect an intention. The AED, then, is computed by averaging the normalized early detection time, $\frac{t}{T}$, over all correctly detected intentions.
Average Correct Duration (ACD): Given an intention observation of length, T, and the total time during which the intention is detected, C, the normalized correct duration is computed as $\frac{C}{T}$. Then, the ACD is computed by averaging over all correctly detected intentions.

5.2. Dataset

For user's intention detection evaluation, we acquire three separate datasets, two of which are acquired using the PR2 in a robotic experimental area and the third is acquired merely in an office using a standalone kinect and smartphone. Their lengths vary between 2700 and 3500 image frames (acquired at 30 fps). The datasets constitute of RGB-D and audio streams. In all cases, the user seated at an approximate

[1] http://cmusphinx.sourceforge.net/

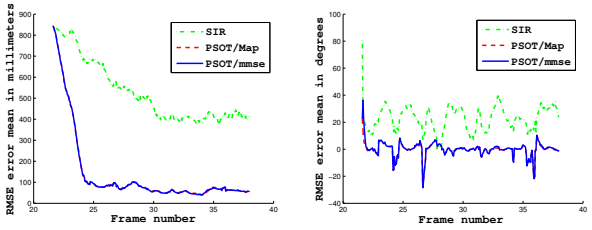
[2] Robot Operating System (ROS): http://wiki.ros.org/

distance of 3m from the RGB-D sensor demonstrates his intention-for-interaction by facing the robot and/or using vocal activity. The datasets are manually annotated to mark intention active regions with the help of the user. The test results presented in section 5.3 are based on one dataset acquired using PR2 (dataset I) and another dataset acquired in an office environment (dataset II). The third dataset, acquired using the PR2, is used to tune and learn the HMM parameters.

To evaluate the PSOT tracker separately, we use a dataset acquired using a kinect and a marker-based motion capture system (Mocap). The Mocap is used to obtain the ground truth for the tracker evaluation and is calibrated and time synchronized with the RGB-D sensor.

5.3. Results and Discussions

Tracker Evaluation The PSOT tracker results obtained using the Mocap coupled dataset are shown in figure 4. For completeness, we have also compared its performance with Sample Important Resampling (SIR) particle filter [15]. As can be seen, the PSOT tracker, either using MAP or MMSE point estimate, outperforms the SIR filter significantly – achieving less than 100 mm average position error and an average angular error centered around 0^o. The results – averaged over 100 runs – are obtained using $\omega = 0.9$, $\psi_p = 0.8$, $\psi_g = 1$, and 100 particles. These parameters are tuned empirically.



(a) Positional accuracy (b) Angular accuracy

Fig. 4. PSOT accuracy evaluation.

Intention Detection This core modality is evaluated using two datasets acquired in robotic and casual office settings. But, initially a separate robotic based dataset is used to learn the different distributions via frequentist approach. Accordingly, $P(z_t^2|x_t)=\begin{bmatrix}0.30 & 0.75\\ 0.70 & 0.25\end{bmatrix}$ rows represent $\{vad, \neg vad\}$ and columns $\{intent, \neg intent\}$. Similarly, the transition matrix, $P(x_t|x_{t-1})=\begin{bmatrix}0.990 & 0.017\\ 0.010 & 0.983\end{bmatrix}$. For $P(z_t^1|x_t)=\mathcal{N}(z_t^1;0,\Sigma)$, Σ is a diagonal matrix with values of 100 (tuned empirically).

Table 1 shows the results obtained for the intention detection modality on the two datasets, I and II. To see the improvement brought by each perceptual component, the evaluation is carried using **VAD** only as measurement, **RGB-D** data input only (PSOT tracker output) as measurement, and the combined **Multimodal** system.

Clearly in all counts, except AED, the proposed multimodal approach outperforms all. In fact, it achieves to detect 72% and 80% of user's intentions correctly with low false alarm rate – 14% and 9% – on dataset I and II respectively. In the robotic dataset, it detects with a 20% lag and manages to flag an intention correctly, on average, over 74% of its sustenance. It also demonstrates quite improved performance on the dataset II. The VAD based approach, though quite fast owing to the high audio frame rate, leads to significant false alarms and less than average TPR on the robotic dataset. This arises because VAD only captures a speech signal without any know how about the intended listener. The RGB-D only approach shows quite promising achievements. The results clearly demonstrate, by fusing a very unreliable measurement like the VAD, which might be overlooked, with RGB-D further perceptual improvements can be gained – in our case a 4% to 8% gain in TPR, a significantly reduced FAR (almost by half in the robotic dataset), and improved correct coverage and early detection.

Figure 5 illustrates an instance taken from dataset I (acquired with the PR2). At this instance, the user turns its attention to the robot and starts talking. Figures 5b and 5c show what the robot sees. The tracked user head pose and shoulder poses are shown in the point cloud depth in figure 5c. The posterior on the user's intention increases in figure 5c flagging this instance as an *intention-for-interaction*. The output of the system for a duration of time is illustrated in figure 6 which corresponds to dataset II (office environment). The figure shows the variation of the posterior over $intent$ and $\neg intent$. Here, a visual correlation could be made between the ground truth annotation (black) and detection output (blue). Both the ground truth and detection outputs take on binary values, but they are shown here scaled to enhance visibility. It is clear that the detection system does well producing results that coincide with the ground truth frequently. Further description of the used dataset and demonstration videos are made available at http://homepages.laas.fr/aamekonn/icme-2015/.

6. CONCLUSIONS

In this work, the importance of multimodal cues in intention-for-interaction detection has been stated. The experiments have shown that audio can significantly improve computer vision and vice versa. Due to omnipotence of context in a robotic application and the importance for intention-for-interaction detection, a further investigation will be to couple our multimodal detector, with an activity detector. This could be justified by the fact that, when a user answer its phone, they usually do not want to start an interaction with the robot. Moreover, context cues could be used to determine any source of interfering noise which would reduce the precision of VAD or any other audio feature.

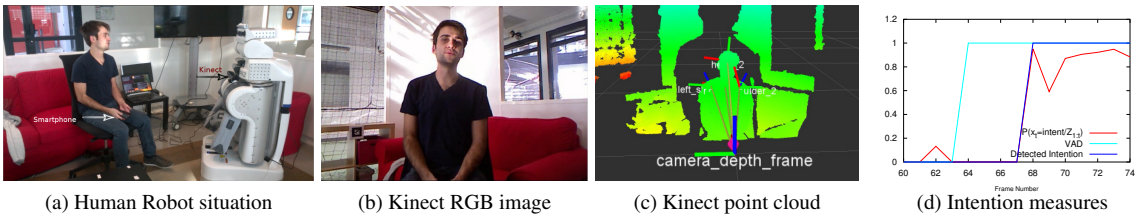
7. REFERENCES

[1] S.-J. Blakemore and J. Decety, "From the perception of action to the understanding of intention," *Nature Reviews Neuroscience*, vol. 2, no. 8, pp. 561–567, 2001.

[2] Y. Xiao, Z. Zhang, A. Beck, J. Yuan, and D. Thalmann,

Table 1. User's intention detection evaluation results on datasets I and II, reported as $\mu(\sigma)$ based on ten repeated runs.

	TPR		FAR		AED		ACD	
	I	II	I	II	I	II	I	II
VAD	0.48 (0.02)	0.56 (0.01)	0.66 (0.04)	0.50 (0.01)	0.01 (0.00)	0.03 (0.06)	0.33 (0.02)	0.56 (0.02)
RGB-D	0.68 (0.05)	0.72 (0.03)	0.26 (0.03)	0.12 (0.03)	0.26 (0.06)	0.10 (0.04)	0.64 (0.12)	0.73 (0.02)
Multimodal	0.72 (0.03)	0.80 (0.02)	0.14 (0.04)	0.09 (0.04)	0.20 (0.08)	0.10 (0.04)	0.74 (0.06)	0.77 (0.03)



(a) Human Robot situation (b) Kinect RGB image (c) Kinect point cloud (d) Intention measures

Fig. 5. Illustrative scene for user intention-for-interaction detection (taken from dataset I).

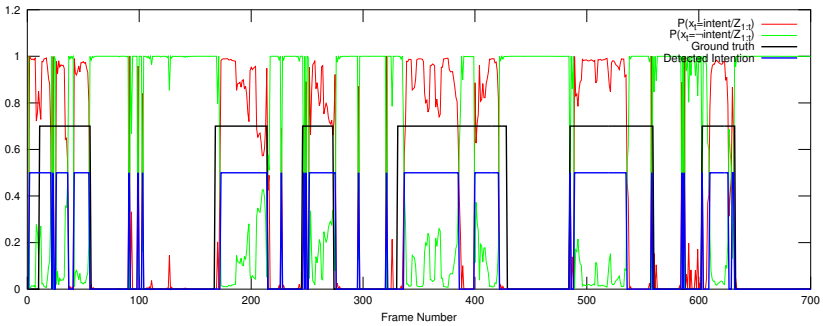


Fig. 6. Sample user's intention detection system output, sequence II, in time showing the posterior, ground truth annotation (in black), and detected intentions (in blue). Ground truth and detection outputs are shown scaled to enhance visibility.

"Human-robot interaction by understanding upper body gestures," *Presence*, vol. 23, no. 2, pp. 133–154, 2014.

[3] B. Huber, "Foot position as indicator of spatial interest at public displays," in *ACM CHI'13 Extended Abstracts on Human Factors in Computing Systems*, Paris, France, 2013.

[4] A. Clair, R. M. St, and M. J. Matarić, "Monitoring and guiding user attention and intention in human-robot interaction," in *ICRA-ICAIR Workshop*, Anchorage, AK, USA, May 2010.

[5] A. Tavakkoli, R. Kelley, C. King, and et al., "A vision-based architecture for intent recognition," in *International Symposium on Visual Computing*, Lake Tahoe, CA, USA, Nov. 2007.

[6] J.-R. Martinez, A. Escobedo, A. Spalanzani, and C. Laugier, "Intention driven human aware navigation for assisted mobility," in *IROS-ASRHE Workshop*, Vilamoura, Portugal, Oct. 2012.

[7] J.-Y. Kuan, T.-H. Huang, and H.-P. Huang, "Human intention estimation method for a new compliant rehabilitation and assistive robot," in *SICE Annual Conference*, Taipei, Taiwan, Aug. 2010.

[8] E. A. Kulić and D. Croft, "Estimating intent for human-robot interaction," in *International Conference on Advanced Robotics*, Coimbra, Portugal, Jul. 2003.

[9] L. Bascetta, G. Ferretti, and P. Rocco et al., "Towards safe human-robot interaction in robotic cells: An approach based on visual tracking and intention estimation," in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, San Francisco, CA, USA, Sept. 2011.

[10] R. Ooko, R. Ishii, and Y. Nakano, "Estimating a users conversational engagement based on head pose information," in *International Conference on Intelligent Virtual Agents*, Reykjavik, Iceland, Sept. 2011, pp. 262–268.

[11] O. C. Schrempf and U. D. Hanebeck, "A generic model for estimating user intentions in human-robot cooperation," in *International Conference on Informatics and Control, Automation, and Robotics*, Barcelona, Spain, Sept. 2005.

[12] G. Fanelli, M. Dantone, J. Gall, A. Fossati, and L. Van Gool, "Random forests for real time 3D face analysis," *International Journal of Computer Vision*, vol. 101, no. 3, pp. 437–458, 2013.

[13] Tim Field, "Openni tracker ROS package (groovy)," http://wiki.ros.org/openni_tracker/, 2013.

[14] J. Kennedy and R. Eberhart, "Particle swarm optimization," in *IEEE International Conference on Neural Networks*, Nov 1995, vol. 4, pp. 1942–1948.

[15] M.S. Arulampalam, S. Maskell, N. Gordon, and T. Clapp, "A tutorial on particle filters for online nonlinear/non-gaussian bayesian tracking," *Transactions on Signal Processing*, vol. 50, no. 2, pp. 174–188, Feb. 2002.