Robust Sensor Fusion for Finding HRI Partners in a Crowd

Shokoofeh Pourmehr, Jack Thomas, Jake Bruce, Jens Wawerla and Richard Vaughan
Autonomy Lab, Simon Fraser University, Burnaby, BC, Canada
spourmeh@sfu.ca

***Abstract*— We present a simple probabilistic framework for multimodal sensor fusion that allows a mobile robot to reliably locate and approach the most promising interaction partner among a group of people, in an uncontrolled environment. Our demonstration integrates three complementary sensor modalities, each of which detects features of nearby people. The output is an occupancy grid approximation of a probability density function over the locations of people that are actively seeking interaction with the robot. We show empirically that simply driving towards the peak of this distribution is sufficient to allow the robot to correctly engage an interested user in a crowd of bystanders.**

I. INTRODUCTION

One long-term aim of autonomous robot research is to have robots work with and around people in their everyday environments, taking instructions via simple, intuitive human-robot interfaces. All else being equal, we would prefer that these interfaces require no special instrumentation of the humans and little or no training. In this paper, we demonstrate such a system, shown in Figure 1 and Figure 2, whereby a self-contained autonomous robot can reliably detect and approach the person in a crowd that most wants to interact with it.

A prerequisite for a successful natural human-robot interaction is for each party to find an interested counterpart. In scenarios with multiple people, the robot must decide which human (if any) to interact with. We want the robot to be able to automatically recognize potentially interested humans present in its workspace and then evaluate the posture, gesture or other salient features of each person to determine their intent to interact.

While studies on attention control typically focus on close range human-robot distances (<2m separation) [1]–[3], mostly on stationary robots, our work looks at controlling a mobile robot's attention in distant (>2m separation) multi-human robot interaction.

This is a challenging task. In addition to ordinary sensor noise, other people may be moving around the environment and occlude the subject; people walking by or performing other tasks will change their appearance to the robot's sensors; the robot's ego-motion changes the sensor readings at every sample; sensor false-positives may mislead the robot. We suggest that there is not a single sensor that can reliably serve.

We achieve robustness by employing an array of multi-modal human detectors and probabilistically fusing their outputs. As a working example, but without loss of generality, we use a laser range finder to detect legs, an RGB camera to detect human torsos, and a microphone array to detect the direction of sound sources. All of these detectors have very different fields of view, detection ranges, and accuracies, while their different modalities allow them to cover each other's weaknesses. The laser, for example, gives us very precise range and bearing measurements, while the microphone array only provides rough directional information. Our fusion method is not limited to these three modalities, but can easily incorporate additional detectors.

To choose sensors and the features they detect, we use our knowledge of simple regularities in human behaviour. For example, among a group of bystanders, a person who is standing facing the robot and calling it will have the highest probability of being a potential interaction partner. We have observed this behaviour combination is generated spontaneously in untrained human subjects [4]. We fuse two independent sources of body pose information with directional audio, placing greater weight on the audio as an actively-generated signal. No single modality is necessary, but we require two modes to agree in order to suppress false

Fig. 1: A live demonstration at HRI'15. The mobile robot is able to robustly track people and approach the most engaging person, despite the noisy and crowded environment.

Fig. 2: Real-world campus setting for experiment IV-C with five uninstrumented users at arbitrary poses. One person, chosen at random, tries to get the robot's attention, and the robot reliably approaches him.

positives. This differs from previous work in active-speaker detection [5]–[7].

The contributions of this paper are: (i) designing a straightforward but effective method for sensor fusion of human detectors that selects the most engaging person to approach for further one-on-one interaction. (ii) demonstrating this method as part of an interaction system for controlling a robot's attention in distant multi-human robot interaction through a series of outdoor experiments. (iii) evaluating this interaction system's performance in a user study with non-expert users. (iv) a ROS-based implementation, freely available online[1], using widely-available sensors.

II. BACKGROUND

To increase the robustness of real-time human detection and tracking, many approaches integrate more than one source of sensory information such as visual and audio cues [5], [8], [9], visual cues and range data [10]–[12] or vision-based and radio-frequency identification (RFID) data [13].

Associating multimodal information with detected humans allows the robot to selectively initiate the interaction with the person with higher interest. Lang et al. [14] proposed a method for a mobile robot to estimate the position of the interaction partner based on 2D laser scanner (leg detection), camera (face detection) and microphone data (sound source location). However, in this system, people have to stand near the robot ($<$ 2m) to be considered as a potential communication partner. Also the user must keep talking to maintain the robot's attention. Our system relates these constraints.

Several authors have worked on enabling a robot to direct its attention to a specific person and/or estimating a user's level of interest in interaction with a robot. Some approaches use distance and spatial relationships as a basis for evaluating engagement. Michalowski et al. [2] and Nabe et al. [3] proposed an approach based on the spatial relationship between a robot and a person to classify the level of engagement. Finke et al. [15] used sonar range data to detect a target person at closer than one meter, based on motion. Muller et al. [16] and Bruce et al. [17] used trajectory information to classify people in the surrounding of the robot as interested in interaction or not. However in some situations having humans approach the robot is infeasible or undesirable, and it is the robot's responsibility to arrive at the target person for one-on-one interaction.

Some work has explored different methods to detect and track multiple speakers [6]. However, our experiments suggest that sound alone does not provide reliable performance in dynamic environments with lots of ambient noise. People can speak, shout or clap to get robot's attention, but by using sound only the robot can get attracted to irrelevant sound sources such as bystanders talking. Okuno et al. [7] developed an auditory and visual multiple-speaker tracking for an upper-torso humanoid robot.

[1] https://github.com/AutonomyLab/autonomy_hri.git

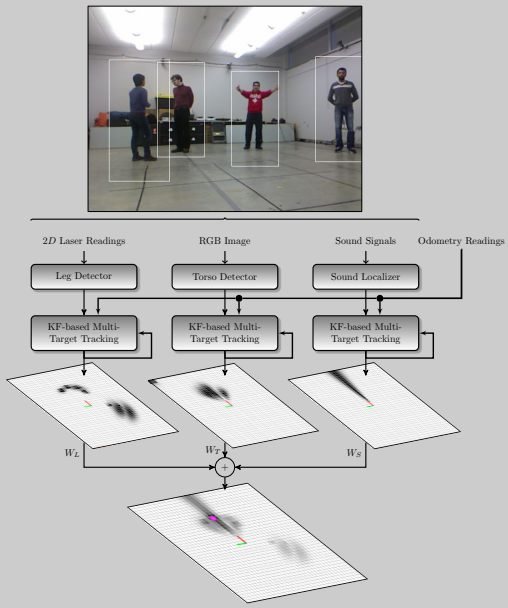

Fig. 3: An overview of the system. Raw sensor data are filtered separately, then projected into evidence grids. Grids are then fused by weighted averaging into a single integrated grid. Grids show real-world data.

In most of these studies, the robot's attention is oriented to the target person by head turning, body turning or eye movements. The person of interest can also lose the robot attention when they stop talking. In this paper we consider a more general situation, where the robot and people are outdoors, mobile, surrounded by distracting people and sound sources, and are in arbitrary locations and poses.

III. SYSTEM DESIGN

A. Multimodal human feature detection

We use a simple probabilistic sensor fusion approach that is easy to understand and implement (Figure 3). The idea of fusing multiple occupancy grids is not novel: Elfes' [18] introduction of the approach showed multi-sensor fusion. Our paper describes the efficacy of this approach for our HRI-partner-finding task. Here, "occupancy" is an estimate of the spatial probability density of finding a partner.

We use three sensors: (i) 2D laser range finder to detect legs, (ii) RGB camera to detect torsos, and (iii) microphone array to detect sound direction. These sensors have different trade-offs in field of view, range and accuracy. They also measure different properties of the user. For example, the leg detector gives accurate location data but is ambiguous about whether the person is facing toward or away from the robot. Sound, on the other hand, is something the user actively emits and is a strong signal for attention-getting, as when calling a dog. As we will explain below, we make explicit use of these differences.

1) Leg Detector: Finding legs in laser range data is a well-explored method for detecting humans. We employed Inscribe Angle Variance (IAV), proposed by Xavier and

Pacheco [19] to find legs by analyzing their geometric characteristics, essentially looking for discontinuities with certain properties in the laser scan. This leg detector runs at 50Hz and provides highly accurate human location information in the robot's coordinate frame with a wide field of view of 270 degrees. A weakness of the leg detector is its high false positive rate. Unfortunately a lot of objects cause similar sensor readings, e.g. furniture, trees, bushes, trash cans.

2) Torso Detector: To detect torsos, we use a camera mounted facing forward at the front of the robot. Grayscale images from the camera are processed to obtain Histograms of Oriented Gradients (HOG) [20] features. These features are robustly classified using linear SVMs trained to detect human torsos. In our system, we use the OpenCV implementation [21] which provides fast multi-scale detections using an image pyramid, and runs at 5Hz on CPU of our mobile-class onboard computer.

To estimate the location of humans, we first compute a bounding box around each torso detection. Given an expected human body size we use the size and image location of the bounding box to estimate the position of a human in the robot coordinate frame. This detector works well at subject distances of up to 10m, however the accuracy is poor in cases of partial occlusions and large deviations of subject height from our median prior.

3) Directional Sound Detector: To detect directional sound we use the Kinect's microphone array. Audio signals are processed using Multiple Signal Classification (MUSIC) [22] to detect the direction of sound sources in the ground plane of the robot frame. We use an implementation of MUSIC from Kyoto University (HARK) [23]. In contrast to the other modalities, the sound detector only provides direction and no true range information for each sound source (since source intensity is unknown). As our goal is to rendezvous, we can use the direction information and rely on the sensor fusion (see below) to obtain position estimates.

Calling the robot by voice, whistle or clap, is a simple and intuitive interface that needs little or no instruction, so we select the loudest detection found by HARK above a certain threshold as an active attention signal. The weakness of sound as an interaction cue is frequent false positives caused by ambient sounds or even echoes. Our system encountered passing buses, talking passers-by and noisy construction equipment. Loud ambient sounds also cause false negatives as the loud signal overwhelms the sensor's ability to detect human voices. We also found that untrained users tend to call the robot occasionally rather than continuously. To reduce the sparsity of sound signals over time, we latch the most-recently-detected sound for two seconds (informally, we observed that this trick was very important for getting good responses to sparse audio).

B. Probabilistic Sensor Fusion Framework

Our proposed framework aims at "multiple-sensor multiple-target" tracking, where human percepts detected through various sensor modalities are all associated with the correct targets. Each of the detectors we have introduced independently tracks different human features and estimates their position relative to the robot frame of reference. The challenge is fusing this data in a way that captures the different characteristics of each sensor, while also being flexible enough to allow new sensors to be added or substituted.

We will address multiple-target tracking at the sensor level through filtering detections into a set of tracks. Afterward, we handle multiple-sensor fusion by converting each sensor's filtered output to a common probabilistic grid format and merging these grids in a weighted average. This adaptive approach allows us to add however many sensors and modalities we want, while also incorporating the characteristics of each sensor through calibrating properties of both the filters and the fusion.

1) Multiple Target Tracking: For each modality, we independently track detected humans using a bank of Kalman Filters (KFs), allowing us to associate evidence collected over time with a particular "track" feature. It also compensates for the robot's motion by incorporating wheel odometry information, but does not model the movement of detected people.

New tracks are spawned when a detection is made beyond a certain threshold distance of any existing track. Otherwise, detections are associated with the nearest neighbouring track. Those tracks that do not receive a measurement update, i.e. no associated detection was made, only have the prediction step of the filter performed - retaining the track but increasing uncertainty. Once a track's uncertainty exceeds a threshold it is removed.

This filtering process provides some robustness against intermittent sensor readings. For example, occlusions, false negatives, and even inconsistent user stimuli (e.g. if a person temporarily stops sending active signals). As each sensor modality has its own filter, it can also have its own thresholds for track association distance and the uncertainty at which it is removed, allowing new modalities to have filters adapted to their particular sensor characteristics.

2) Probabilistic Grids: The middle-step that allows us to fuse the results from different sensor modalities is converting the output of the Kalman Filters into probabilistic evidence grids. These grids are similar to occupancy grids [18] but instead of holding the probability of an obstacle, we store the probability that an attentive subject is at each location.

For this, we compute a location probability distribution for each tracked human feature using a modality-specific sensor model. In our implementation, leg detections are modelled with a normal distribution. For torso detection, we use a multi-variant normal distribution to reflect the fact that range estimates are not very reliable. Sound detections are modelled using a cone along the measured direction vector. This is a simple model of the likely distribution of ranges of a user who is calling the robot. The probability distribution for each modality is then discretized into a separate evidence grid.

3) Sensor Fusion: To compute the integrated probability distribution for all detected humans, a fused evidence grid is calculated as the weighted average of corresponding grid

cells from all other grids. Each modality-specific grid is centered over the robot, ensuring that detections from the same human will overlap. Example grids are shown in Figure 3. The integration weights for each modality are assigned based on sensor characteristics and uncertainties.

We have some a priori reasoning in our implementation for choosing the *relative* weights: since sound is actively generated it may be more likely to indicate interest, while legs and torsos are possessed by interested and uninterested people alike. Hence, we assigned the highest weight to the (S)ound evidence grid. In our experience, the (T)orso detector exhibits fewer false positives than the (L)eg detector, so we assigned a higher weight to the torso grid than the leg grid. This results in an implicit ordering from most-reliable combinations to least-reliable combinations of [TLS], [TS], [LS], [TL]. This means for example that if two people were calling out, and both had their legs detected, but only one had a visible torso, we would prefer the person with a visible torso since that person would probably be facing the robot and would thus be directing her attention to it.

Detections made by only one sensor modality (e.g. [T], [L], or [S]) are treated as inherently unreliable, to avoid detection errors such as the legs of a chair or distant ambient noise. For this reason, the cone modelling sound direction is capped at 10m, as our other two modalities cannot reach further out.

The advantages of fusing sensor data after tracking rather than before include being more modular, allowing sensor modalities to be added or subtracted from the system. It allows our final evidence grid to represent all of the sensor-specific tuning of the tracking filters, probability distributions and weighting. It also handles correlating detections from different locations across the body of the robot and with different fields of view, so long as each sensor's grid has one transformation back to the center of the robot for the fused result.

C. Attention Control and Behaviour Design

The integrated evidence grid can now be used to generate the robot's behaviour. Several methods could be considered, but we chose a very simple and explicable approach to demonstrate the efficacy of the sensor fusion: we simply find the highest probability in the evidence grid and servo the robot towards that location. As the robot moves the evidence grid is continuously updated and the robot corrects the approach vector. This enables the user to move and be followed by the robot and it gives the robot an opportunity to recover from false sensor readings. Once the robot has approached the human to within 2 meters the robot stops. To give the impression that it is ready for a close range interaction it plays a happy sound. If the person does not respond, the robot gives up, plays a sad sound and turns away looking for another person.

If all probabilities in the evidence grid are below a given threshold its detections are considered unreliable. In this case, the robot turns to sweep its sensors over the environment to find humans.

The user and the robot form a tight interaction loop that appears similar to that between a dog and its owner. By observing the robot, the user can deduce if the robot is paying attention to her (approaching) or not. If the robot is not paying attention the user can simply provide more stimuli, e.g. call louder or orient more towards the robot.

IV. EXPERIMENTAL RESULTS

We implemented the designed system on a typical outdoor mobile robot, Husky by Clearpath Robotics. The robot is equipped with a Kinect providing the RGB camera and a four microphones array, and a 2D SICK laser scanner. The sensors have different but overlapping fields of view. Legs can be detected in a 270 degree arc up to a distance of 10 meters. The camera has a 57 degree horizontal FOV and is capable of detecting human torsos at distances up to 8 meters. The microphone array has a detection zone of 180 degrees in front of the robot but only reports bearing and not range.

Three experiments were performed to validate our probabilistic sensor fusion framework's ability to select an interaction partner. Afterward, an evaluative user study was conducted with non-expert users to assess the system's performance in an HRI scenario. In all four experiments, the robot is co-located with a group of people including one who wants to initiate an interaction (the *interactor*). This person will stand facing the robot and occasionally call for it verbally.

A. Experiment A: Framework only

The first experiment is designed to test the reliability of the sensor fusion by selecting the most promising person seeking robot's attention in an artificial setting. It is not our intended HRI scenario, but rather an exhaustive test of the system's functionality by exposing it to a wide range of possible detections at once. The robot's objective is to pick the interactor from a group of 8 research assistants 7m away. Subjects are positioned outdoors in a semi-circle with a 7m radius around the robot and approximately 2 meters apart from each other.

We systematically set up distractions by positioning people in a way that each shows a different subset of attractive features. For example, we ask some to cover their legs, some to stay quiet, and some to stand outside the camera/torso-detector field of view. Only the preferred interactor presents the full set of legs, torso and occasional sound to the robot.

The robot is given a 10 second time window to determine the location of the interactor. Actually approaching the selected human for interaction is omitted here in order to focus on the reliability of the attention system.

We call a selection successful if the robot "favours" the interactor during this period. We define favour to mean that the detected interactor position is closest to the true position of the interactor for longer than it is closer to any of the distractors.

Users take turns taking the role of interactor and varying their appearance to the robot according to a predefined script

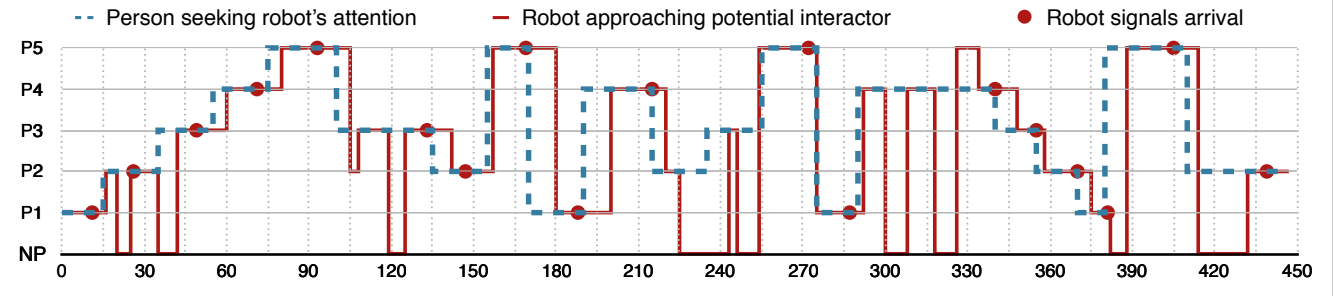

Fig. 4: Results from experiment IV-C: Diagram of the robot's responses to rapidly switching the interactive role between five people (P1-P5) at random. The blue dashed line marks the time line of which subject is seeking attention, the red solid line shows which person the robot is paying attention to and the red dots indicate when the robot entered close range interaction state. The robot usually attends to the correct person.

ensuring all permutations were tested. The robot correctly identified the right person on 21 out of 24 trials (87.5%), giving 99% confidence this approach improves on selecting one detected person at random. Failures occurred when ambient sound was coming from the same direction as a distractor person, whose legs and torso were detected (our test location had loud intermittent construction noise in the background).

B. Experiment B: Testing discrimination at range

We placed two research assistants outdoors at a distance of 7 meters in front of the robot and varied the distance between the people in order to test how well the sensor fusion system could discriminate between adjacent humans. The robot was now allowed to approach a detected interactor to also examine interference from motion and changing distances. We measured the success rate and time required for the robot to reach the correct target, where a trial was successful if the robot was facing the correct person when it stopped. Results of 65 trials (5 repeats for each distance) are presented in Figure 6.

In trials where the people are standing very close to each other (<1.5 meters), the system has difficulty distinguishing the individual humans. This is mainly due to the relatively large uncertainty in the sound source direction detection. In these cases, the robot approached the centre between the 2 people. For strictness, we declared these outcomes as failures, but for most practical purposes the correct person is now within close interaction range.

At each distance there were some cases where the robot was distracted by the non-interactor participant but recovered when the interactor kept calling the robot. The further apart the two humans were, the more off-course this "wavering" could pull the robot and thus the higher average arrival time and variance. At 12 meters or more, the participants were at the extreme range of our sensors, making initial detection difficult and sometimes drawing the robot too far away to recover.

C. Experiment C: Playing tag with five people

In the third experiment, we examined the robustness and responsiveness of the system in a dynamic environment

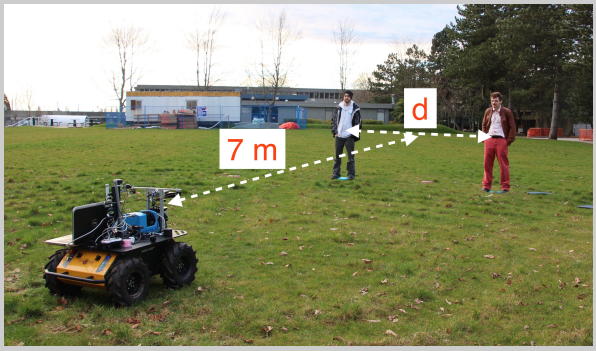

Fig. 5: Experiment IV-B: Two people stand 7 meters in front of the robot. One person seeks the robot's attention. We empirically determine the minimum distance *d* between the people at which the robot can no longer distinguish the attentive user from the bystander.

Fig. 6: Experiment IV-B: Success rate and approach time in relation to distance between subjects.

where the role of interactor would switch over the course of one continuous test. We instructed five research assistants to stand in arbitrary positions surrounding the robot (see Figure 2).

One person was selected at random to be the first interactor. The interactor stands still and calls the robot in a normal voice, while the other research assistants walk around the vicinity of the robot as bystanders. The robot approaches the strongest fused detection, and when the robot stops directly in front of its chosen person it plays a "happy sound" to

indicate its readiness to engage in a one-on-one interaction. If this person is the interactor, she moves away and chooses a new interactor at random. If she is not, she ignores the robot, which times-out and returns to scanning for new interactors. A section of this experiment is shown in the accompanying video[2].

The timeline of interactions is shown in Figure 4, plotting the time when each of five people (P1-P5) were in the interactor role, and the time when the robot was focused on them or on no-person (NP), and the moment (dots) when the robot correctly announced it was ready for a one-on-one.

In seven and a half minutes, the robot engaged in 20 interactions. In 18 cases, the robot successfully found the interactor and correctly announced its arrival. However, we observed that in two cases between 220 and 260 seconds, the robot would find the target for a short time, but become distracted by another person and did not find the correct interactor.

D. Experiment D: System Performance with Non-Expert Users

We evaluated the effectiveness of the proposed system in a detailed user study with a sample of the university community in a semi-controlled indoor setting. The study recruited 34 participants (23 females, 11 males), ranging in age from 17 to 73, with the majority being 20 to 30 years old. They were asked to call the robot over from across an 8×5m room without leaving a fixed spot, with other humans nearby as bystanders. The participant and three research assistants stood at the far end of the room, with the participants and one assistant facing the robot and two other assistants facing each other and conversing (Figure 7).

The system was fully explained beforehand. This means telling them the robot will look for human legs, body and direction of sound to detect the potential interaction partner and works best if they stand facing the robot and call it to get its attention.

In $\%85$ of trials the robot successfully distinguished the participant as the interested person. Of the five failure cases, all were related to audio detection and the behaviour of the participant. Two participants clapped as their chosen audio signal but did so too infrequently to be detected, so that the robot was drawn to the two chatting bystanders instead. In two more cases, the participants stopped trying to call the robot once it was approaching them, allowing the robot to be distracted by the chatting bystanders again. The last failure case had the robot approach the silent bystander possibly as a result of misidentified echo from the adjacent participant.

V. DISCUSSION AND FUTURE WORK

The experiments used to both validate and evaluate our system focused on raw performance but do not address improvements to the human-robot interaction experience. In the user-study, fourth experiment, we recorded other performance and user preference data that will be presented elsewhere.

[2]https://youtu.be/KtAz_fJUGmo

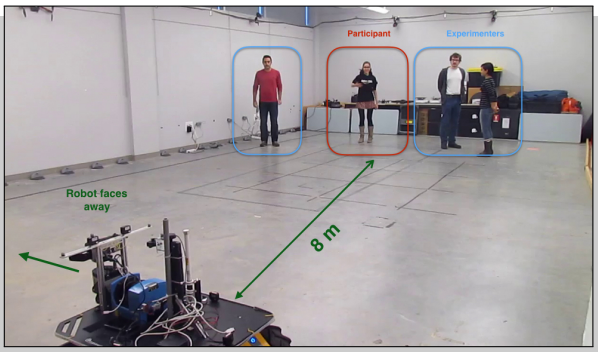

Fig. 7: Experiment IV-D: Study Setup

But the data presented in this paper and [4] provide evidence that inexperienced users humans were able to reliably obtain the robot's attention and call it over for interaction using only their instinctive behaviour and without any instrumentation (i.e. they carried no special equipment or clothing). This is specifically important for robots deployed in public settings, as untrained users can engage in an interaction or call the robot's attention with little or no instruction. Failure cases from each experiment also suggests some possible improvements. Differentiating between human-sourced audio like words and clapping versus environment noise might usefully improve the reliability of sound as a detection method. Speech recognition might also be useful in distinguishing between active encouragement and discouragement signals.

The Husky robot platform is adapted for outdoor use, and proved somewhat unsuitable to indoor social environments due to its size, appearance and movements. We showed that the same system works indoors and out, but our indoor work will use a telepresence robot form factor in future.

This work considers human-robot interaction over relatively large distances compared to almost all the literature ($>$2m separation). We noticed but have not yet exploited the way people interact with the robot varies over the course of an interaction as their mutual distance changes (we omit evidence here for lack of space). We expect that the robot's behaviour and sensing should also change with mutual distance.

Environmental factors can also modify human behaviour, where the intensity of the interaction signals may increase with the intensity of the social setting - a loud party might cause users to call loudly, or to prefer gestures to calls in a library setting. Adapting sensor fusion parameters to the current setting could be a useful extension.

VI. CONCLUSIONS

We proposed a system which integrates detected human features from multiple modalities for a mobile robot to choose the most likely person interested in a close interaction in a robot-multi-human scenario. Our probabilistic sensor fusion framework combined passive and active stimuli to successfully direct the robot's attention. A series of real-world experiments in outdoor uncontrolled environments, a user study with dozens of non-expert participants, and a live demo

at HRI '15 in a crowd of hundreds of people all demonstrate the practicality of our approach. Our ROS implementation is freely available online (`goo.gl/TFV8y5`).

References

[1] V. Chu, K. Bullard, and A. Thomaz, "Multimodal real-time contingency detection for hri," in *Intelligent Robots and Systems (IROS 2014), 2014 IEEE/RSJ International Conference on*, Sept 2014, pp. 3327–3332.

[2] M. Michalowski, S. Sabanovic, and R. Simmons, "A spatial model of engagement for a social robot," in *Advanced Motion Control, 2006. 9th IEEE International Workshop on*, 2006, pp. 762–767.

[3] S. Nabe, T. Kanda, K. Hiraki, H. Ishiguro, K. Kogure, and N. Hagita, "Analysis of human behavior to a communication robot in an open field," in *Proceedings of the 1st ACM SIGCHI/SIGART Conference on Human-robot Interaction*, ser. HRI '06. New York, NY, USA: ACM, 2006, pp. 234–241. [Online]. Available: http://doi.acm.org/10.1145/1121241.1121282

[4] S. Pourmehr, J. Thomas, and R. Vaughan, "What untrained people do when asked make the robot come to you (late-breaking abstract)," in *Proceedings of the 2016 ACM/IEEE International Conference on Human-robot Interaction*, ser. HRI '16, 2016.

[5] H.-J. Bhme, T. Wilhelm, J. Key, C. Schauer, C. Schrter, H.-M. Gro, and T. Hempel, "An approach to multi-modal humanmachine interaction for intelligent service robots," *Robotics and Autonomous Systems*, vol. 44, no. 1, pp. 83 – 96, 2003, best Papers of the Eurobot '01 Workshop. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0921889003000125

[6] M. Murase, S. Yamamoto, J.-M. Valin, K. Nakadai, K. Yamada, K. Komatani, T. Ogata, and H. G. Okuno, "Multiple moving speaker tracking by microphone array on mobile robot." in *INTERSPEECH*. ISCA, 2005, pp. 249–252.

[7] H. Okuno, K. Nakadai, K. Hidai, H. Mizoguchi, and H. Kitano, "Human-robot interaction through real-time auditory and visual multiple-talker tracking," in *Intelligent Robots and Systems, 2001. Proceedings. 2001 IEEE/RSJ International Conference on*, vol. 3, 2001, pp. 1402–1409 vol.3.

[8] B. Kühn, B. Schauerte, K. Kroschel, and R. Stiefelhagen, "Multimodal saliency-based attention: A lazy robot's approach," in *Proc. 25th Int. Conf. Intelligent Robots and Systems (IROS)*, Vilamoura, Algarve, Portugal, October 7-12 2012.

[9] K. P. Tee, R. Yan, Y. Chua, Z. Huang, and S. Liemhetcharat, "Gesture-based attention direction for a telepresence robot: Design and experimental study," in *Intelligent Robots and Systems (IROS 2014), 2014 IEEE/RSJ International Conference on*, Sept 2014, pp. 4090–4095.

[10] P. Poschmann, S. Hellbach, and H.-J. Bhme, "Multi-modal people tracking for an awareness behavior of an interactive tour-guide robot," in *Intelligent Robotics and Applications*, ser. Lecture Notes in Computer Science, C.-Y. Su, S. Rakheja, and H. Liu, Eds., vol. 7507. Springer Berlin Heidelberg, 2012, pp. 666–675.

[11] N. Bellotto and H. Hu, "Multi sensor-based human detection and tracking for mobile service robots," *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on*, vol. 39, no. 1, pp. 167–181, Feb 2009.

[12] C. Martin, E. Schaffernicht, A. Scheidig, and H.-M. Gross, "Multi-modal sensor fusion using a probabilistic aggregation scheme for people detection and tracking," *Robotics and Autonomous Systems*, vol. 54, no. 9, pp. 721 – 728, 2006.

[13] T. Germa, F. Lerasle, N. Ouadah, V. Cadenat, and M. Devy, "Vision and rfid-based person tracking in crowds from a mobile robot," in *Intelligent Robots and Systems, 2009. IROS 2009. IEEE/RSJ International Conference on*, Oct 2009, pp. 5591–5596.

[14] S. Lang, M. Kleinehagenbrock, S. Hohenner, J. Fritsch, G. A. Fink, and G. Sagerer, "Providing the basis for human-robot-interaction: A multi-modal attention system for a mobile robot," in *in Proc. Int. Conf. on Multimodal Interfaces*. ACM, 2003, pp. 28–35.

[15] M. Finke, K. L. Koay, K. Dautenhahn, C. L. Nehaniv, M. L. Walters, and J. Saunders, "Hey, I'm over here - How can a robot attract people's attention?" in *IEEE International Symposium on Robot and Human Interactive Communication*, 2005.

[16] S. Muller, S. Hellbach, E. Schaffernicht, A. Ober, A. Scheidig, and H.-M. Gross, "Whom to talk to? Estimating user interest from movement trajectories," in *IEEE International Symposium on Robot and Human Interactive Communication*, 2008.

[17] J. Bruce, J. Wawerla, and R. Vaughan, "Human-robot rendezvous by co-operative trajectory signals," in *Proc. 10th ACM/IEEE International Conference on Human-Robot Interaction Workshop on Human-Robot Teaming*, 2015.

[18] A. Elfes, "Occupancy grids: A stochastic spatial representation for active robot perception," in *Autonomous Mobile Robots: Perception, Mapping, and Navigation (Vol. 1)*, S. S. Iyengar and A. Elfes, Eds. Los Alamitos, CA: IEEE Computer Society Press, 1991, pp. 60–70.

[19] J. Xavier, M. Pacheco, D. Castro, A. Ruano, and U. Nunes, "Fast line, arc/circle and leg detection from laser scan data in a player driver," in *Robotics and Automation, 2005. ICRA 2005. Proceedings of the 2005 IEEE International Conference on*, April 2005, pp. 3930–3935.

[20] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Compp uter Society Conference on*, vol. 1. IEEE, 2005, pp. 886–893.

[21] G. Bradski, "OpenCV: the open source computer vision library," *Dr. Dobb's Journal of Software Tools*, 2000.

[22] R. Schmidt, "Multiple emitter location and signal parameter estimation," *Antennas and Propagation, IEEE Transactions on*, vol. 34, no. 3, pp. 276–280, Mar. 1986. [Online]. Available: http://dx.doi.org/10.1109/TAP.1986.1143830

[23] K. Nakadai, H. Okuno, H. Nakajima, Y. Hasegawa, and H. Tsujino, "An open source software system for robot audition hark and its evaluation," in *Humanoid Robots, 2008. Humanoids 2008. 8th IEEE-RAS International Conference on*, Dec 2008, pp. 561–566.