A Sensor Fusion Framework for Finding an HRI Partner in Crowd

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

***Abstract*— We present a probabilistic framework for multi-modal information fusion to address the detection of the most promising interaction partner among a group of people, in an uncontrolled environment. To achieve robust operation, the system integrates three multi-modal percepts of humans and regulates robot' behaviour to approach the location with the highest probability of an engaged human. A series of real-world experiments demonstrates the robustness and practicality of the proposed system for controlling robot's attention.**

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 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 Fig. 1 and Fig. 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 of a successful natural human-robot interaction is for each party to find a counterpart for interaction. In scenarios with multiple people, the robot must decide which human (if any) to interact with. Here, we want the robot to be able to automatically recognize the 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) [1]–[3], mostly on stationary robots, our work looks at controlling a mobile robot's attention in distant 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 target; 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: e.g. a picture of a human on the wall may attract the robot's attention but should not cause it to wait for an interaction indefinitely. 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, a 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. The laser, for example, gives us a very precise range and angle 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.

Fig. 1: The mobile robot is able to robustly track people and attend the most engaging person to deliver its service, despite the noisy and crowded environment. (live demonstration at HRI'14)

To address the problem of approaching the potential interaction partner, among a group of people, we incorporated auditory cues, as an active stimulus, with different modality cues of human presence. We assume if a person is standing facing the robot, and calling it, among a group of people, he or she is probably the most interested person in having an interaction. This particularly differs from talking person detection, since even if the human doesn't call the robot, it can still navigate to the location of detected people, one at the time. However, in order to draw the robot's attention, the user should send more information through an active communication signal.

The contributions of this paper are: (i) designing an interaction system for controlling the robot's attention in distant multi-human robot interaction. (ii) demonstration of a simple but effective method for sensor fusion of human detectors that selects the most engaging person to approach for further one-on-one interaction; (iii) a ROS-based implementation, freely available online, using widely-available sensors. We demonstrate its effectiveness in real-world outdoor robot experiments.

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

Fig. 2: Real-world setting (university campus) for experiment IV-A 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.

[4]–[6], visual cues and range data [7]–[9] or vision-based and Radio-frequency identification (RFID) data [10].

Associating multi-modal information with detected humans allows the robot to selectively initiate the interaction with the person with higher interest. Lang et al. [11] 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 should keep talking to maintain the robot's attention.

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. [12] used sonar range data to detect a target person at closer than one meter, based on motion. Muller et al. [13] and Bruce et al. [14] 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 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 [15]. However, our experiment suggests that sound alone does not provide reliable performance in dynamic environments with 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. Okuno et al. [16] developed an auditory and visual multiple-speaker tracking for an upper-torso humanoid robot.

In most of these studies, the robot's attention is oriented to the target person by head turning, body turning or eye movements. Also the person of interest can loose the robot attention when 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. In these situations, it is hard to find the correct interaction partner among the crowds of people. Therefore, we propose a system in which

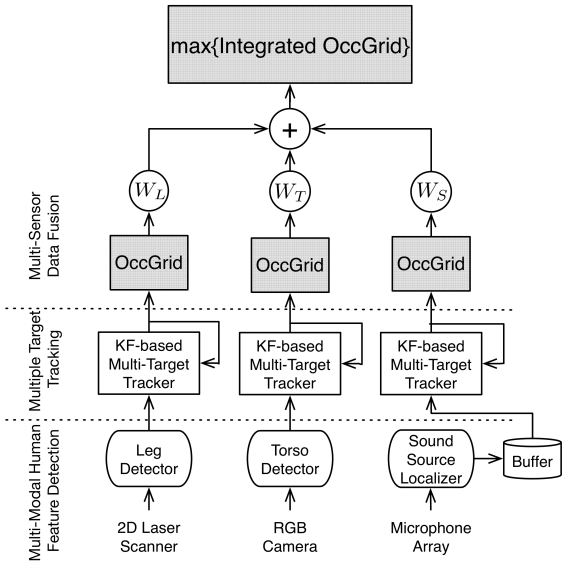

Fig. 3: An overview of the system showing how different components are connected. (*OccGrid* = Occupancy Grid), (*KF* = Kalman Filter)

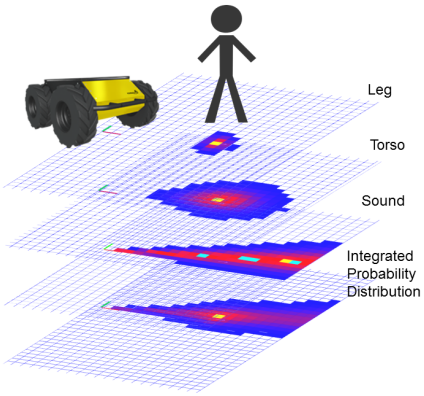

Fig. 4: Evidence grids use to fuse detections for each sensor modality separately (top three grids). Grids are then fused by weighted averaging into an integrated grid (bottom grid).

the robot is able to adaptively change its focus, approach and initiate a close interaction with the person with currently highest apparent interest.

Our goal is to design an interaction system that is robust, reliable and can be deployed in public settings. A series of real-world experiments in outdoor, uncontrolled environments (university campus), with up to 8 human participants, and a live demonstration at HRI '15 in a crowd of hundreds of people, demonstrates the practicality of our approach.

III. SYSTEM DESIGN

A. *Multi-modal human feature detection*

We use a simple probabilistic sensor fusion approach that is easy to understand and implement. An overview of the system is shown in Fig. 3. The approach of fusing multiple occupancy grids is not novel [17] : this paper serves as a case study and demonstrates the sufficiency of this approach for this task. Our implementation can easily be adapted and

extended for similar problems and different sensors and robot hardware.

Here we use three sensors: (i) laser range finder to detect legs; (ii) camera to detect torsos and (iii) a microphone array to detect sound with direction. These sensors have different trade-offs in field of view, range and accuracy. They also measure different properties of the user, each with different information about the intent to interact. 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, a thing the user actively emits, on the other hand, is a strong signal of 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-used method for detecting humans. We employed Inscribe Angle Variance (IAV), proposed by Xavier and Pacheco [18] to find legs by analyzing their geometric characteristics. A leg detection directly provides a human location in the robot's coordinate frame. This detector runs at 50Hz and provides highly accurate location information with a wide field of view of 270 degrees. The downside of the leg detector is a high false positive rate. The detector essentially looks for discontinuities with the right geometric properties in the laser scan. Unfortunately a lot of objects cause similar sensor readings, e.g. furniture, trees, bushes, trash cans, etc.

2) Torso Detector: To detect torsos, we use the Microsoft Kinect RGB camera mounted looking forward at the front of the robot. The camera has a horizontal field of view of 57 degrees. Grayscale images from the camera are processed to obtain Histograms of Oriented Gradients (HOG) [19] features. These features are robustly classified using linear SVMs trained to detect human torsos. In our system, we use the OpenCV implementation [20] which provides fast multi-scale detections using an image pyramid, and runs at 5Hz on CPU on 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 outputs at 5Hz and works well at subject distances of up to 10m. The directional accuracy is good, but the range accuracy is poor in the 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 microphone array of the Kinect. Audio signals are processed using Multiple Signal Classification (MUSIC) [21] to detect the direction of a sound in the ground plane of the robot frame. We use an implementation of MUSIC from Kyoto University (HARK) [22]. In contrast to the other detectors, the sound detector only provides direction and no range information. In principle, it would be possible to move the robot to a different vantage point (i.e. drive perpendicular to the sound direction) and then triangulate the location of the sound source. But this would be time-consuming and cause the robot to exhibit an unusual search behaviour. Since our goal is the rendezvous, we can simply 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. The weakness of sound as an interaction cue is frequent false positives caused by ambient sounds. 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 found that 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 2 seconds (informally, we observed that this trick was very important for getting good responses to sparse audio).

B. Multiple Target Tracking

Each of the detectors independently detects one or more human features and estimates their position relative to the robot frame of reference. For robustness, we accumulate evidence of each detection over time while taking the robot's motion into account. It is, therefore, important that we accurately track each detection before fusing the different modalities into a unified detection.

For each modality, we independently track each human feature using a bank of Kalman Filters (KFs). We empirically tuned the measurement model of each sensor to reflect their particular behaviour including uncertainty. The motion of the robot is estimated using wheel odometry and is used in the process model of the Kalman filter. The motion of individual people, however, is not explicitly modelled.

To associate a detection with a track, we use the nearest neighbour. A new track is spawned if the distance to the closest neighbour exceeds a threshold. If a track did not receive a measurement update, i.e. no associated detection was made, only the prediction step of the KF is performed. Consequently, the tracks are retained but the uncertainty increases. Once the uncertainty exceeds a threshold the track is removed. By choosing separate thresholds for each sensor modality, we can tune the system's respond to specific sensor characteristics.

This provides two benefits, (i) it enables the system to handle intermittent sensor data, for example due to temporary occlusions and false negatives; and (ii) the user need not provide continuous stimuli. The latter is important for the system to feel natural, for example calling the robot once, then wait for a reaction and possibly call again is a more natural and less strenuous interaction compared to non-stop calling.

C. Multi-Sensor Data Fusion

In the previous step, we obtained a set of Kalman filters tracking detections independently for each modality. Next we have to fuse these into a unified estimate of human attention seeking so we can control the behaviour of the robot.

In a first step, we compute a probabilistic evidence grid for each sensor modality. These grids are similar to occupancy grids [17] but instead of holding the probability of an obstacle, we store the probability of a human seeking attention. For this, we compute a location probability distribution for each detection 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. And sound detections are modelled using a cone along the measured direction vector. The cone length is limited to 10 meters. The probability distribution for each modality is then discretized into a separate evidence grid.

To compute the integrated probability distribution, a fourth evidence grid is calculated as the weighted average of corresponding grid cells in all modality-specific evidence grids. Example grids are shown in Fig. 4.

The integration weights are assigned to each modality based on sensor characteristics and uncertainties. We have some a priori reasoning for choosing the relative weights: since sound is actively generated it may be more likely to indicate interest while legs and torsos are apparent in 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 of TLS, TS, LS, TL. This means for example that if two people are calling out, and both have their legs detected, but only one has a visible torso, we prefer the person with visible torso since that person is probably facing the robot and is thus directing her attention to it.

D. Attention Control and Behaviour Design

The integrated evidence grid can now be use to generate behaviour and create a natural, easy to use and reliable interaction between the user and the robot. For this, we 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 values in the evidence grid are below a given threshold the robot observed no human or only unreliable detections. In this case, the robot randomly turns and searches for humans until it finds one. We define detections made by only one sensor modality as unreliable, e.g. leg detections without a torso detection are often caused by furniture and not by people.

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 easily 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. Informally, we observed that this interface feels very natural.

IV. EXPERIMENTAL RESULTS

We performed three different experiments in an outdoor uncontrolled environment (university campus). We implemented the designed system on a typical mobile robot, Husky by Clearpath Robotics. The robot is equipped with a Kinect providing the RGB Camera and a 4 channel microphone array, and a 2D SICK laser scanner. All these 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 60 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.

In all following experiments, the robot is co-located with a group of people including one who wants to initiate an interaction. This person will stand facing the robot and occasionally call for it verbally.

A. Experiment A: Playing tag with five people

In the first experiment, we examined the robustness and responsiveness of the system in a dynamic environment. We instructed 5 people to stand in arbitrary positions surrounding the robot (see Figure 2). One person was selected "at random" to be the person who seeks the robot's attention (the *interactor*). The interactor stands still and calls the robot in a normal voice. The robot approaches the strongest fused detector response. When the robot stops directly in front of its chosen person it plays a "happy sound" to indicate its readiness to engage in the one-on-one interaction. If this person is the interactor, she moves away and chooses a new interactor at random. If the chosen person is not the interactor, she ignores the robot, which times-out and returns to scanning for new interactors. This process continued for 8 minutes. A section of this experiment is shown in the accompanying video.

In eight minutes, the robot managed to correctly locate and engage the interactor 24 times. The timeline of interactions is shown in Figure 5, plotting the time when each of five people (P1-P5) were in the interactor role, and 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 19 cases, the robot successfully found the interactor correctly first try and correctly announced this. The robot also recovered from false positives and negatives in most cases. However, we observed that in some cases, the robot found the target for a short time, but got distracted by another person (between 220 and 260 seconds). In addition, in one case the robot approached the interactor correctly, but did not announce its arrival. This happened at 460 seconds, where the dot is missing.

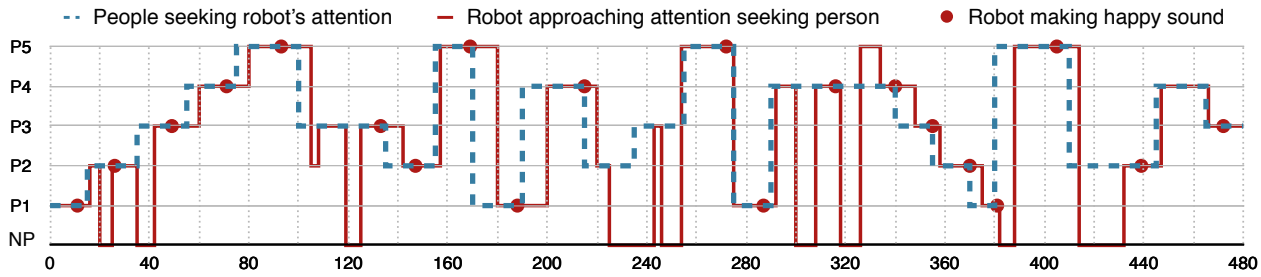

Fig. 5: Results from experiment IV-A: Diagram of the robot 's responses to rapidly switching the interactive role between five people (P1-P5) at random. The blue dashed line marks the time line of which subject is seeking attention, the red solid line shows which person the robot is paying attention to and the red dots indicate when the robot entered close range interaction state.

B. Experiment B: Detection only

This experiment is designed to test that the sensor fusion works correctly to select the most promising interaction in an artificial setting. It is not our intended HRI setting. The robot's objective is to pick from a group of 8 people and distractors 7m away only the person who seeks the robot's attention. Subjects are positioned in a semi-circle with 7 meter radius around the robot and approximately 2 meters apart from each other (see Figure 6.).

We systematically setup distractions by positioning people in a way that each shows a different subset of the 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 interactor presents legs, torso and occasional sound to the robot.

The robot is given 10 second time window to determine the location of the interactor. The approach phase is omitted here because we hereby want to investigate the reliability of the attention system only.

We call a selection successful if the robot "favours" the interactor during this period. We define favour to mean that the highest probability of the integrated evidence grid should be closest to the true position of the interactor for a longer time than any other stimulus.

The human subjects take turns taking the role of interactor and varying their appearance to the robot according to a predefined script ensuring all permutations were tested. The robot correctly identified the right person on 21 out of 24 trials (%87.5) with %99 confidence compared to randomly selecting one person among all detected people. 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 construction noise in the background). Also, if the robot did not pick the right person immediately we labelled the trial a failure.

C. Experiment C: Testing discrimination at range

In the third scenario, we placed two people at 7 meters distance in front of the robot and varied the distance between the people. We measured the success rate and time required for the robot to reach the correct target. If the robot stopped facing the correct person we labeled the trial a success.

Fig. 6: Setup of experiment IV-B: Eight human participants are positioned in a semi-circle with radius 7 meter around the robot. Individuals create specific sensor stimuli by shouting, covering their legs or standing outside the field of view of a particular sensor.

Results of 65 trials (5 repeats for each distance) are presented in Fig. 8.

In the trials where the people are standing very close to each other (1 meter and 1.5 meters), the system has difficulties distinguishing the individual humans. This is mainly due to the relatively large uncertainty in the sound source direction detection.

In this case, 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. In some cases the robot was sometimes distracted by the other person but recovered when the interested person keep calling the robot. This wavering increased the time to arrive at the interactor when the distance between the people was high. We observed that when the distance is more than 8 meters, the right person always gets robot's attention but the approach is simply longer due to the symmetry and it takes more time.

When the distance were larger than 12 meters, the two people were at the extreme range of any of our sensors, so the robot could not immediately detect people and pick the right target. In this case, it had to wander around looking for people which explained the lower success rate and more travel time.

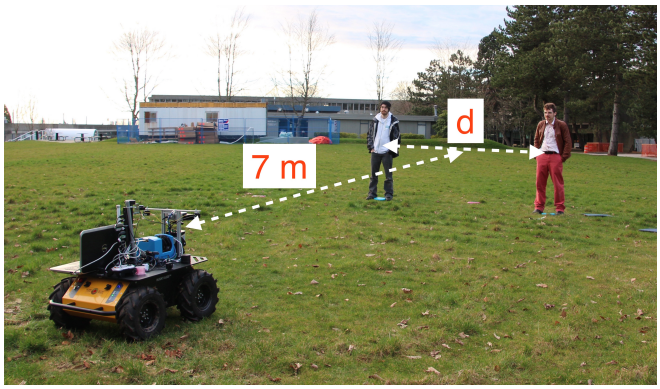

Fig. 7: Experiment IV-C: Two people stand 7 meters in front of the robot with varying distance between each other. One person seeks robot's attention.

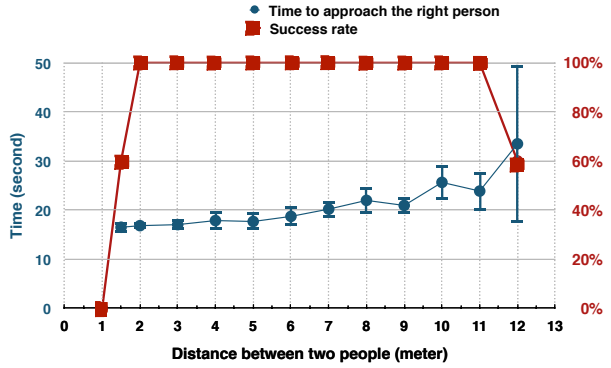

Fig. 8: Experiment IV-C: Success rate and approach time in relation to distance between subjects.

V. DISCUSSION AND FUTURE WORK

The users who participated in evaluation experiments and in the live demonstration at the HRI'14 conference are part of the author's research group. In this section, we briefly reflect our observations and experiences during these trials. In the future, we plan to evaluate the system usability, user experience and social acceptance in a detailed user study with the general public, in an extreme uncontrolled social setting.

We informally claim that this designed system uses simple and robust methods for deploying robots in crowded environments. People can use natural and intuitive communication signals to interact with the robot and attain its attention whilst easily understand its behaviour. This is specifically important for robots deployed in public settings, as untrained and non-technical users can engage in an interaction or call the robot's attention, with no special instrumentation of the humans and little or no instruction.

Despite the intuitive interaction system, the shape and appearance of our mobile robot was not tailored for indoor social environments (e.g. conference hall or cocktail parties). However, its platform is designed for outdoor applications, e.g. ground search and rescue, where the proposed system can be used in the task of finding and approaching the people who need robot's service. We believe, even in a case of two robots with the same types of sensors and interaction system, the form factor of a robot affects people's social perception of it.

In addition to people's reactions to the form and structure of the robot, the real world environment conditions influence the human behaviour. As we observed, the intensity of interaction increases with the intensity of the social setting. In crowded places with lots of people talking to each other, the level of ambient noise and false positives is very high. In these situations the interactor has to make greater effort in getting and maintaining the robot's attention, which may affect their patience and motivation.

Also, as one the objective of this work is regulating distant multi-human-robot interaction (distances >3m), we noticed that the way the interested person acts differ depending on the distance from the robot. In future work, we plan to evaluate and quantify the impact of environment properties including crowd size and relative human-robot distance of the people's experience in interacting with the robot, using the proposed interaction system, and subsequent system performance.

VI. CONCLUSIONS AND FUTURE WORK

We have demonstrated a system which integrates detected human features from three modalities for a mobile robot to choose the person who is more likely interested in having close interaction in a robot-multi-human scenario. We showed that combining passive and active stimuli can be used to designate a particular person among a population for subsequent one-on-one interactions. A series of real-world experiments in outdoor uncontrolled environments (university campus) with up to 8 human participants, and a live demonstration at HRI '15 in a crowd of hundreds of people, demonstrates the practicality of our approach.

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] 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

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[14] 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 Conference on Human-Robot Interaction Workshop on Human-Robot Teaming*, 2015.

[15] 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.

[16] 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.

[17] 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.

[18] 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.

[19] 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.

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

[21] 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

[22] 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.