SOCIALLY ACCEPTABLE APPROACH TO HUMANS BY A MOBILE ROBOT

Mercedes Garcia-Salguero, Javier Monroy, Alejandro Solano and Javier Gonzalez-Jimenez
Machine Perception & Intelligent Robotics group (MAPIR). Dept. of System Engineering and Automation.
Biomedical Research Institute of Malaga (IBIMA). University of Malaga. Spain.
mercedesgarsal@uma.es

ABSTRACT

In this paper we propose a solution to endow a mobile robot with the ability to approach humans in a safe and socially acceptable way. Our proposal focuses on real world indoor environments where the usual presence of multiple humans and obstacles notably rise the complexity of the approach action. We first deal with the problem of accurately estimating the 3D poses of all humans in the work space (positions and orientations), to then focus on the estimation of the most appropriate navigation goal from which the robot should start the interaction with the user. For the latter, we define a cost function that, accounting for multiple proxemic parameters and complying with the restrictions inherent of a robot navigating in a human environment, enables ascertaining an optimal solution. Different experiments are presented to demonstrate the feasibility of our proposal to work under real world conditions.

CCS CONCEPTS

• **Computing methodologies** → **Robotic planning**; **Motion path planning**; **Vision for robotics**;

KEYWORDS

Human Pose Estimation, Human-Robot Interaction, Computer Vision, Mobile Robot, People Detector, OpenPose, ROS, Proxemic, People Approach

ACM Reference format:
Mercedes Garcia-Salguero, Javier Monroy, Alejandro Solano and Javier Gonzalez-Jimenez . 2019. SOCIALLY ACCEPTABLE APPROACH TO HUMANS BY A MOBILE ROBOT. In *Proceedings of 2nd International Conference on Applications of Intelligent Systems, Las Palmas de Gran Canaria, Spain, January 7–9, 2019 (APPIS 2019)*, 7 pages.
https://doi.org/10.1145/3309772.3309793

1 INTRODUCTION

Human-Robot Interaction (HRI) is a multidisciplinary field which has been developed during years, achieving excellent results in many domains [5]. Recent growth in personal robots and their cohabitation with humans encourage researchers to focus on new applications like assisting in the location of lost or odorous objects [17, 18, 22], or the integration of social behaviours into them [3]. Characteristic human skills like understanding daily or spontaneous conversations, or being able to interpret facial expressions of emotion (nonverbal communication) are key tasks for robots in order to achieve a friendly and "social" behaviour while interacting with people. Several projects have implemented successfully some of these aspects, for instance, initiating a conversation with a person, understanding human gestures or following people [4, 25].

The physical presence of a service robot among humans, jointly with the necessity of establishing communication with them leads to tackle the problem of how robots may approach humans as other people would do. This problem involves two major tasks: (i) the detection and pose estimation of the humans present in the environment, and (ii) the planning of the navigation path to approach the target user in a socially acceptable way, both for the target user as well as for the rest of humans that may be present in the scene.

The first of these tasks, human identification, is a fundamental step in order to approach a person successfully. Human body complexity and diversity of motions provoke that both detection and pose estimation become difficult tasks to solve, specially under the challenging conditions a social mobile robot must operate: limited on-board resources, real-time constraints and real-world environments. In this context, we find works that have addressed this problem from different perspectives and considering a wide variety of sensors. Multiple works have addressed human detection employing only RGB images (see [24] for a review), obtaining excellent results with the introduction of deep neural networks. To list a few, works like [19] propose efficient solutions able to overtake the complexity associated with human movements and partial occlusions in the scene. Exploiting the depth information provided by modern RGB-D cameras, recent works started to combine colour images with depth data to increase the accuracy of the human detection [21].

Nevertheless, the aforementioned works do not compute the actual human pose in the 3D space (something mandatory to our problem), but focus on detecting people in the image frame. Inferring the 3D human pose is not a straightforward problem given that the complexity of the human body generates a high-dimensional probabilistic tracking problem, computationally expensive to solve. Notwithstanding, tentative solutions have been proposed based on monocular images, multi-camera setups to infer depth data, or 3D range sensors [1, 23]. The main limitation when bringing these solutions to a service mobile robot is the limited number of sensors

This work has been partially funded by the European Union under H2020-ICT project MoveCare (732158) and by the Science and Innovation Ministry of Spain under project WISER (DPI2017-84827-R).

APPIS 2019, January 7–9, 2019, Las Palmas de Gran Canaria, Spain

ACM ISBN 978-1-4503-6085-2/19/01…$15.00
https://doi.org/10.1145/3309772.3309793

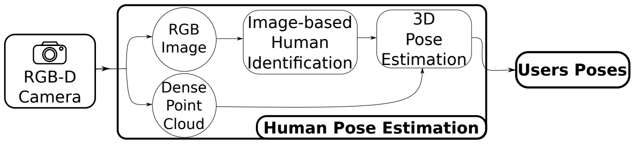

Figure 1: Flow diagram of the human detection process. For every captured image a real-time multi-person system detects, locates and processes all humans in the image frame. Then, exploiting depth information, the 3D poses of the users are estimated.

and computational resources on-board, as well as the real-time constraint to be socially acceptable. In this work we propose a minimal solution that only requires a RGB-D camera, implementing a detection model based on neural networks to identify humans in the RGB image, and estimating their 3D poses (position and orientation) by exploiting the depth data.

Once the humans in the environment have been detected and their poses have been estimated, we need to face the problem of social navigation and user approach. Due to the different types of human-related interactions and their social background, this task is far from trivial. Particularly, the study of human use of space and its effect on behaviour, communication, and social interaction -known as *proxemic* [6]- provides several "rules" to model this social behaviour, *e.g.* how far apart individuals engaged in conversation stand depending on the degree of intimacy between them, or what is the minimum distance to a person that must be respected when passing close to him/her without intention to establish a conversation. Study and application of proxemic distances during human-robot interaction has been studied previously, *e.g.* [15], proposing algorithms that, for instance, allow robots to move around people in a "*natural*" and safe way, to deliver objects to the users from an appropriate distance or to approach people in order to initiate a conversation without intimidating them [3, 11]. Yet, most of these works focus on simplified scenarios where either the detection of the user is simulated or given by an external detection system (*i.e.* a distributed sensor network), or they consider working environments with wide open areas where the robot is free to move complying with the proxemic restrictions (*e.g.* a mall).

In this paper we seek to provide a mobile service robot with the ability to approach a human in a safe and socially acceptable way. This involves estimating the pose (position and orientation) of the different humans in the environment and the planning of a safe, optimum and socially acceptable path toward the target user in order to start the HRI. Our goal is to achieve this behaviour relying only on a 2D laser rangefinder (just for navigation purposes [8]) and a 3D RGB-D camera, both mounted on the service robot. Furthermore, we are interested in practical situations where the robot has to work under the challenging conditions of real-word environments, *e.g.* with the presence of obstacles as tables, desks, chairs, etc, the presence of other humans, or the inherent complexity of multi-room scenarios.

2 HUMAN POSE ESTIMATION

A fundamental skill to endow a service robot with the ability to interact with humans is that of user detection and robust pose estimation. The objective is not only to estimate the pose of the

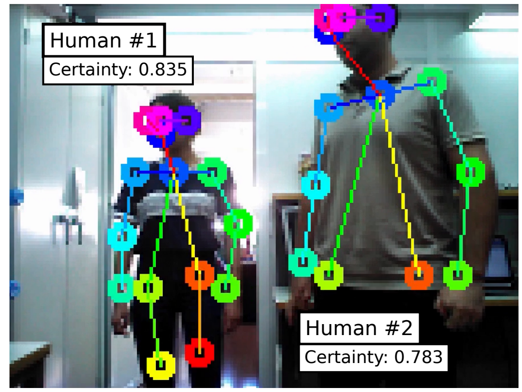

Figure 2: Image captured by the robot and illustration of the detected body joints and links, as well as the overall certainty score for each detected human.

target user (the one we are interested in starting an interaction with), but the pose of all humans in the scenario that may influence the navigation path. In this work we propose a human pose estimation algorithm that, relying only on the rich information provided by a RGB-D camera, estimates the poses of all humans in the scene by considering two consecutive stages (see Figure 1).

2.1 Image-based Human Identification

Human 2D pose estimation is an extensively studied problem. Different and varied approaches have been proposed to estimate the location and orientation withing the image frame of multiple users [1]. Exceptionally relevant are the recent techniques that address this identification task by deploying sophisticated and efficient convolutional neuronal networks (CNN) [2, 26]. CNN are able to identify multiple users on a single colour image in a real time basis even under real world conditions (*i.e.* considering cluttered environments and the presence of multiple textures in the background).

In this work we employ *OpenPose* [2], an efficient method for multi-person 2D pose estimation. Concretely, we work with the version trained with the COCO data set [12] and depth-wise convolution[1]. *OpenPose* efficiently infers the human poses in the input image, returning a set of keypoints or "body parts" (composing a "skeleton"), along with an individual score for each keypoint detected. With the only purpose of improving the performance of this phase we migrate the model, originally implemented in the *Caffe* deep learning framework [9], to *TensorFlow*[2], an open source software library for high performance numerical computation.

Then, to strengthen the identification and remove false positives, we post-process the output -skeletons- by estimating an overall certainty value $\Gamma \in [0, 1]$ for each detected human in the image. Figure 2 shows an example of this process by superimposing to the original RGB image the human "skeleton" provided by *OpenPose* together with the computed certainty value. The latter is computed by averaging the different individual scores for each body part. This parameter allows purging poor identifications by setting a minimum threshold (currently a parameter set empirically to $\Gamma = 0.5$). Γ is of special interest for the case of mobile robots where this human

[1]https://github.com/ildoonet/tf-pose-estimation
[2]https://www.tensorflow.org

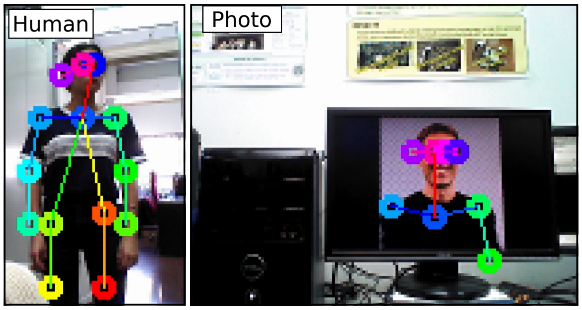

Figure 3: Illustration of false-positive human detection when relying only on RGB data. In both scenes a human is detected and its body parts are estimated: (left) Image of a real human standing in front of the robot, (right) Image captured by the robot of a human picture on a TV. Our depth-based filter efficiently removes these false-positives.

detection process needs to be executed with the limited resources on-board. Γ helps to select the N best detected candidates in the image, notably reducing the computational cost of the subsequent processing stages (see Section 2.2).

2.2 3D Pose Estimation

To compute the 3D poses of the humans referring to the world reference frame we fuse the skeletons detected in the colour image with the depth information from the RGB-D camera (in the form of a dense point cloud of the scene). The objective is to estimate the location and orientation of the different humans -constituted by a set of 3D points (X, Y, Z) known as 3D body parts-. The following sections describe in detail the depth computation and the orientation estimation, as well as additional filtering processes to avoid false positives not detected in the previous stages.

2.2.1 Position Estimation. To estimate the entire 3D human body position we compute the different 3D body parts which composed it. We map the pixels of the colour image with the indices of the dense point cloud (registration), directly enabling the computation of the 3D body parts from the previously computed 2D ones. Furthermore, the 3D centre of the human body is computed by averaging the detected body parts, and the variance among the corresponding depth values is calculated. The latter is used to discriminate if the detected human is indeed a real human or, on the contrary, it's a human photography, a picture, a painting, or an image on the TV, among other likely scenarios to be faced by a service robot [20]. It must be noticed that image-based solutions (Section 2.1) incorrectly identifies those cases as humans as exemplified in Figure 3. This discrimination (filtering stage) is easily achieved by forcing a minimum value on the depth-variance associated to each candidate human.

2.2.2 Orientation Estimation. The last stage for obtaining the 3D pose of the humans in the scene is the estimation of their orientations. This is a crucial step for conveniently enabling a robot to approach a user. Only a correct estimation of the user's orientation allows the robot to approach the human in a social acceptable way, and furthermore, to look towards him/her prior starting the interaction. Different approaches have been proposed to estimate the user orientation *e.g.* focusing on the detection of the users' shoulders [1]. Yet, either the high computational requirements of some solutions or the unreliability of the others due to the common occlusions of body parts in the image, lead to most of these solutions not be feasible in practical applications.

In this work we opt for simplifying this estimation by considering only four possible human orientations: *looking right*, *looking toward the robot*, *looking left* and *looking backward*, referring always to the robot reference system. To do so, we employ the representative features of the human head, namely, eyes and ears. In particular, we work with 4 body parts from the set of keypoints which are returned by the human detector. Human orientation is inferred attending to the presence or not of these body-parts in the image, *e.g.* if just the left ear is detected, we can assume that the right one is hidden -usually behind the head- implying the human is heading left. Importantly, if none of the 4 key features appear within the image, but a human has been detected (getting through the different filters previously described), we conclude that this particular human is heading backward to the robot.

3 PATH PLANNING

The second task to be executed when approaching a user (once the 3D poses of all humans in the environment have been estimated), is to generate an optimum, socially acceptable and safe path toward the target human. In this work we assume that the robot is able to identify the target human from all other possible people in the work space, focusing on the definition of the proxemic areas and the estimation of the navigation goal to approach it.

According to Hall [6], during a normal conversation, humans maintain a distance between 46 cm to 120 cm (aka "personal and social spaces") which depends on social factors such as gender, culture, or degree of friendship among other factors. Closer distances are considered "intimate space" to be always avoided, while farther ones fall in the so called "public space" where we can assume the user perceives no intention to start an interaction. Exploiting this valuable information we tune the parameters of the *multi-layered costmaps* approach [13] to define a proxemic area (cost function used for navigation) for each detected human as a 2D symmetric Gaussian function centred around the user location. This Gaussian spreads up to a maximum distance of 120cm, entailing that a robot will try to avoid invading the proxemic areas (aka personal and social spaces) of all users in the environment except for the one it wants to interact with. One of the advantages of employing the *layered costmaps* approach is that semantically-separated layers are combined into a master layer to be used in the robot path planning phase. Each particular layer tracks one type of obstacle or constraint. In this work, we stack four different layers accounting for the floor-plan of the environment, the obstacles detected by the robot's on-board sensors, their inflations and the above described proxemic layer.

Last, but not least, we need to determine the navigation goal for the robot to fulfil the approach. This navigation goal must fall within the proxemic distances associated to human interaction, that

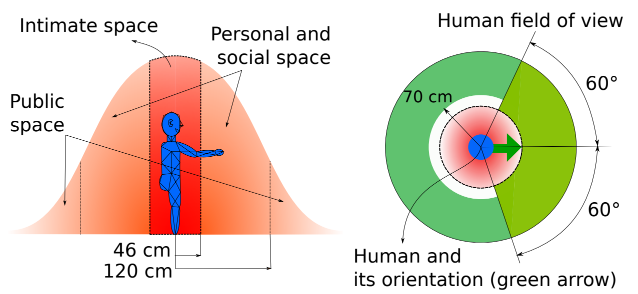

Figure 4: (left) Implementation of the human proxemic areas as a 2D Gaussian function centred at the user location. (right) Top view of the human proxemic area and detail of the regions defined when calculating the most convenient navigation goal to approach that user (light and dark green).

is, between 46 to 120 cm from the target user. Furthermore, to establish a comfortable interaction, it is desirable that this navigation goal falls within the field of view (FOV) of the target human [16], avoiding shocks or turns that may negatively influence the posterior interaction. Taking this into account, the restrictions that apply (from a proxemic point of view) when determining the navigation goal are depicted in Figure 4, considering a human FOV of approximately 120° [10]. It must be stressed that, in real environments, it might be the case where no valid goal can be set within the FOV of the target user, in such cases we restrict even more the admissible area by rising the minimum distance to the user from 46 to 70 cm.

To get the optimal navigation goal $p_g = (x_g, y_g, \theta_g)$ with respect the target human pose $p_t = (x_t, y_t, \theta_t)$, we define the cost function $\Psi \in \mathbb{R}$ accounting for four different variables:

- $d(p_g, p_t) \in \mathbb{R}$, the Euclidean distance from the target human pose p_t to the goal robot position p_g, favouring those navigation goals that get closer to the user.
- $\alpha(p_g, p_t) \in \mathbb{R}$, the angle between the robot final pose heading θ_g and the human heading θ_t, promoting those goals that fall in front of the user.
- $c(p_0, p_g) \in \mathbb{R}$, the overall travelled distance by the robot to reach the goal pose p_g from its initial pose $p_0 = (x_0, y_0, \theta_0)$, preferring short distances in order to lessen the navigation time.
- $\Phi(p_g, p_t) \in \{0, 1\}$, a boolean stating if p_g falls withing the FOV of the target human, penalising those poses which do not fulfil this condition.

Likewise, we define four free parameters $(w_d, w_\alpha, w_{path}, w_{fov}) \in \mathbb{R}$ to weight the aforementioned variables, and which allow us to determine the value of Ψ for a candidate goal pose p_g as follows:

$$\Psi = w_d d(p_g, p_t) + w_\alpha \alpha(p_g, p_t) + w_{path} c(p_0, p_g) + w_{fov} \Phi(p_g, p_t)$$

According to the specific values of $(w_d, w_\alpha, w_{path}, w_{fov})$ (see Section 4 for a detailed description of the influence of each parameter on the estimation of the navigation goal), we search for the pose $p_i = (x_i, y_i, \theta_i)$ that minimises Ψ. It is important to notice that in this optimisation problem we take into account the following constraints: (i) the goal must be free of obstacles, (ii) it must be reachable by the robot, that is, a valid navigation path must exists from the current robot pose to the target goal, (iii) the goal must fall within the user's personal or social spaces ($d(p_g, p_t) \in [46, 120](cm)$) and (iv), when feasible, it is preferable to approach the user within its FOV.

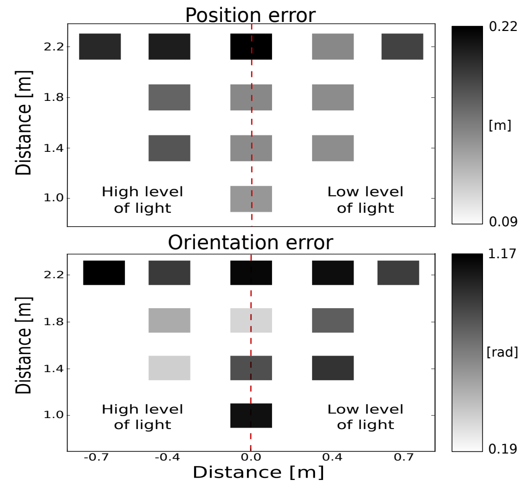

Figure 5: Average errors in the user's pose estimation (averaging more than 200 repetitions) for different relative poses between the robot (situated at (0,0)) and the human.

4 EXPERIMENTS AND RESULTS

In this section we present three experiments designed to validate the proposed approach. The first one is designed to measure the error in the user's pose estimation, both position and orientation, while the second experiment seeks to assess the behaviour of the system under challenging conditions (*e.g.* when the target user is sitting behind a table or heading towards a wall) and to analyse in detail the influence of the model parameters when calculating the most convenient navigation goal to approach an user. Finally, the last experiment, aimed to demonstrate the system feasibility to work in real world scenarios, is carried out in an office-like environment with a Giraff mobile robot [14].

4.1 Pose Error Estimation

The errors associated to the user's detected pose can be separated into position and orientation (see Section 2.2). The former accounts for the post-processing of the depth data while the latter is related to the accuracy in the human body detection in the RGB image. Figure 5 shows average values of both errors for different relative poses between the user and the robot. Moreover, two light conditions have been tested to measure the influence of this important parameter. As can be seen, position errors increase with distance and are relatively affected by light conditions, showing more error when working under highly illuminated environments. This is related to the infrared pattern projected by RGB-D cameras when estimating depth, which is affected by distance and natural light.

Estimation of the user's orientation is performed only with RGB data and its based on the detected human body-parts. Close distances lead to less body parts being detected (as the user doesn't fit in the camera FOV), while too far distances reduce the human size in the image frame, degrading the orientation estimation. Finally, and in contrast to position estimation, natural light does not interfere

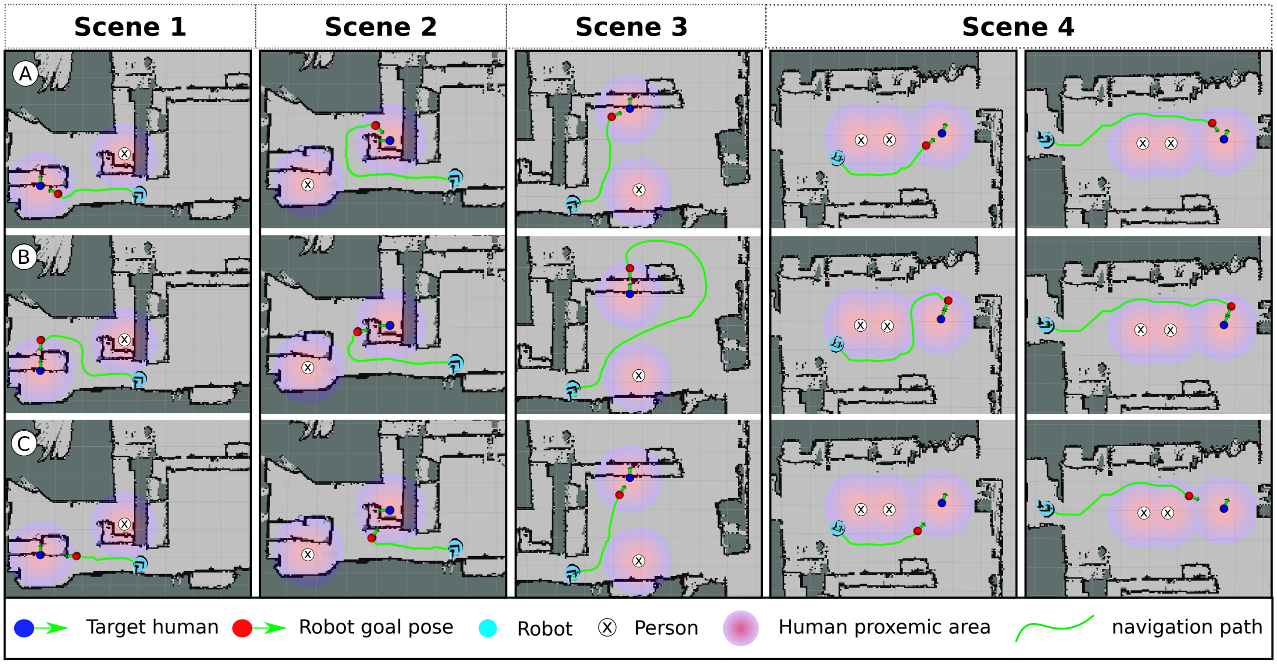

Figure 6: Navigation paths computed in four different scenes with multiple humans present and distinct initial poses of the robot. The robot navigation goal varies withing the same scene depending on the parameters configuration

in the orientation process but improves the image quality, yielding slightly better results. Yet, orientation errors are considerably high, suggesting that more elaborated methods must be considered, for example by also accounting for depth data.

4.2 Validation Experiment

In this experiment we test our approach in a multi-room environment endowed with the usual office furniture (*e.g.* tables, chairs, bookshelves, etc) and the presence of multiple humans. The objective is to analyse how the different parameters of the proposed cost function Ψ influence the selection of the navigation goal when approaching the user. Therefore, in this experiment we simulate the user detection, manually setting the user location and orientation in order to evaluate only the path planning component. Concretely, we evaluate three different parameter configurations: *(A)* emphasising the importance to get close to the user, that is, rising the distance weight parameter w_d, *(B)* minimising the angle constraint to promote goals that fall in front of the user (increasing the value of w_α), and *(C)* favouring short navigation paths by enhancing the w_{path} factor. In all cases, we consider a constant value for the cost associated to the user's FOV, $w_{fov} = 0.4$.

Furthermore, we test the behaviour of the system under four challenging conditions:

1. **Scene 1**: The target user is sitting behind a rectangular table.
2. **Scene 2**: The target user is sitting behind a L-shape table, being more restrictive the available navigation goals.
3. **Scene 3**: The target user is facing a wall (*e.g.* the common case where a person is sitting in front of a table which is besides a wall that delimits the room).
4. **Scene 4**: The target user is in a free-space room but two other people maintaining a conversation are also present in it. In this case, we consider two different initial poses of the robot to illustrate the robustness of the proposed solution.

Figure 6 depicts the approaching-path followed by the robot attending to the different scenes and parameter configurations. For each case, we indicate the initial robot pose, the location and orientation of the users simulated in the environment (with their corresponding proxemic areas), the navigation goal selected after evaluating the cost function Ψ and the path planned by the robot to reach it. As can be noticed, different parameter configurations lead to completely different goal poses, particularly on those scenes with presence of obstacles inside the proxemic area of the user. These results illustrate the importance of properly setting the weight parameters according to the user's social background and, although less important, according to the environment configuration. For example, configuration *(A)* favours goal poses which are closer to the user even if other possibilities would have allowed the robot to start the interaction from inside the user's FOV (*e.g.* scenes 1-A and 4-A). On the other hand, putting too much emphasis on those poses in front of the user, as in configuration *(B)*, may lead to erroneous approaches like the one shown in scene 3-B where the robot set the navigation goal in a different room. The latter can be easily amended by rising the weight associated to the navigation distance w_{path} (see scene 3-C for example). Finally, it must be noticed the capacity of the proposed system to deal with the presence of multiple people in the work space, complying with the proxemic areas of all detected humans when navigating toward the target user.

4.3 Real Experiment

Finally, to evaluate the complete system (from user detection to path planning toward the selected navigation goal) and to test the performance under real world conditions, we carry out a real experiment. In this case we use a Giraff mobile robot [14], equipped with an Orbbec Astra RGB-D camera for people pose estimation, and a 2D Hokuyo URG laser rangefinder for navigation purposes [7]. Additionally, a Nvidia Jetson-TX2 board has been mounted on the Giraff robot to achieve real-time computation.

The experiment flow is as follows: (i) The robot is initially commanded to navigate to a random position within the work space, (ii) then it starts turning on itself in order to ascertain the presence of humans around it (stopping as soon as a positive detection is obtained). (iii) Then it evaluates the cost function to determine the most appropriate navigation goal, and (iv) executes it. Results of this experiment can be seen in a compilation video hosted in:

http://mapir.uma.es/work/appro_user.

This video shows multiple scenes where the robot approaches users in diverse circumstances, that is, varying the initial pose of the robot as well a the number of humans and obstacles present. We demonstrate that our approach is able to adapt to these challenging situations and successfully complete the approaching action.

5 CONCLUSIONS AND FUTURE WORK

In this work we have revised the problem of social navigation by a mobile robot and focus on the critical task of approaching the user safely and in a socially acceptable way in order to start an interaction. Specifically, our proposal have been designed for real world environments where the presence of obstacles (both static and dynamic) have an important impact in the final navigation path.

We have analysed the two main tasks involved in the social approach process, namely, user pose estimation and assessment of the most convenient navigation goal. For the former we have employed a solution based on convolutional neuronal networks and improve it by fusing depth information to both, filter-out false positive identifications and to properly estimate the user 3D pose (position and orientation). Then, we have proposed a cost function that, attending to a set of parameters related to the user proxemic areas, allows the declaration of the optimal navigation goal to approach a target user.

The system has been experimentally evaluated, making strong emphasis on the influence of the model parameters when estimating the navigation goal, as well as to assess the performance when dealing with complex situations as users sitting behind tables or surrounded by other humans. Furthermore, a compilation video is provided with several scenes where a real mobile robot approaches a human in an office-like environment with the usual furniture and multi-room configuration of these environments.

As future work we envisage two different lines of action. On the one hand, to improve the human detection accuracy (and 3D pose estimation) by introducing a motion model of the people involved which may allow us to track them and to infer next moves. On the other hand, related to the navigation goal to approach the user, to analyse how the four free parameters of our optimisation algorithm vary according to social features such as culture, traditions or human gender among others. In this sense, an extensive experiment needs to be carried out in order to allow the robot learn the preferences of different users.

REFERENCES

[1] Drazen Bršcic, Takayuki Kanda, Tetsushi Ikeda, and Takahiro Miyashita. 2013. Person tracking in large public spaces using 3-D range sensors. *IEEE Transactions on Human-Machine Systems* 43, 6 (2013), 522–534.

[2] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. 2016. Realtime multi-person 2D pose estimation using part affinity fields. *arXiv preprint arXiv:1611.08050* (2016).

[3] Terrence Fong, Illah Nourbakhsh, and Kerstin Dautenhahn. 2003. A survey of socially interactive robots. *Robotics and autonomous systems* 42, 3-4 (2003), 143–166.

[4] Rachel Gockley, Jodi Forlizzi, and Reid Simmons. 2007. Natural person-following behavior for social robots. In *Proceedings of the ACM/IEEE international conference on Human-robot interaction*. ACM, 17–24.

[5] Michael A Goodrich, Alan C Schultz, et al. 2008. Human–robot interaction: a survey. *Foundations and Trends in Human-Computer Interaction* 1, 3 (2008), 203–275.

[6] Edward T Hall, Ray L Birdwhistell, Bernhard Bock, Paul Bohannan, A Richard Diebold Jr, Marshall Durbin, Munro S Edmonson, JL Fischer, Dell Hymes, Solon T Kimball, et al. 1968. Proxemics [and comments and replies]. *Current anthropology* 9, 2/3 (1968), 83–108.

[7] Mariano Jaimez, Javier Monroy, and Javier Gonzalez-Jimenez. 2016. Planar Odometry from a Radial Laser Scanner. A Range Flow-based Approach. In *IEEE International Conference on Robotics and Automation (ICRA)*. 4479–4485.

[8] M. Jaimez, J. Monroy, M. Lopez-Antequera, and J. Gonzalez-Jimenez. 2018. Robust Planar Odometry Based on Symmetric Range Flow and Multiscan Alignment. *IEEE Transactions on Robotics* (2018), 1–13.

[9] Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell. 2014. Caffe: Convolutional Architecture for Fast Feature Embedding. *arXiv preprint arXiv:1408.5093* (2014).

[10] Kiyoshi Kiyokawa. 2007. A wide field-of-view head mounted projective display using hyperbolic half-silvered mirrors. In *Proceedings of the 2007 6th IEEE and ACM International Symposium on Mixed and Augmented Reality*. IEEE Computer Society, 1–4.

[11] Kheng Lee Koay, Dag Sverre Syrdal, Mohammadreza Ashgari-Oskoei, Michael L Walters, and Kerstin Dautenhahn. 2014. Social roles and baseline proxemic preferences for a domestic service robot. *International Journal of Social Robotics* 6, 4 (2014), 469–488.

[12] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*. Springer, 740–755.

[13] David V Lu, Dave Hershberger, and William D Smart. 2014. Layered costmaps for context-sensitive navigation. In *Intelligent Robots and Systems (IROS 2014), 2014 IEEE/RSJ International Conference on*. IEEE, 709–715.

[14] Matteo Luperto, Javier Monroy, Francisco-Angel Moreno, J.R. Ruiz-Sarmiento, Nicola Basilico, Javier Gonzalez-Jimenez, and N. Alberto Borghese. 2018. A Multi-Actor Framework Centered around an Assistive Mobile Robot for Elderly People Living Alone. In *Robots for Assisted Living Workshop - International Conference on Intelligent Robots and Systems (IROS)*.

[15] Ross Mead, Amin Atrash, and Maja J Matarić. 2013. Automated proxemic feature extraction and behavior recognition: Applications in human-robot interaction. *International Journal of Social Robotics* 5, 3 (2013), 367–378.

[16] Ross Mead and Maja J Matarić. 2016. Perceptual models of human-robot proxemics. In *Experimental Robotics*. Springer, 261–276.

[17] Javier Monroy and Javier Gonzalez-Jimenez. 2017. Gas Classification in Motion: An Experimental Analysis. *Sensors&Actuators: B. Chemical* 240 (2017), 1205–1215.

[18] Javier Monroy, J. R. Ruiz-Sarmiento, Francisco-Angel Moreno, Francisco Melendez-Fernandez, Cipriano Galindo, and Javier Gonzalez-Jimenez. 2018. A Semantic-Based Gas Source Localization with a Mobile Robot Combining Vision and Chemical Sensing. *Sensors* 18, 12 (2018).

[19] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*. 91–99.

[20] JR Ruiz-Sarmiento, Cipriano Galindo, and Javier Gonzalez. 2011. *Improving human face detection through ToF cameras for ambient intelligence applications*. Springer, 125–132.

[21] Yang Shi, Xiucheng Dong, Decai Shi, and Qiuyan Yang. 2016. Human Detection Using Color and Depth Information by Kinect Based on the Fusion Method of Decision Template. *ICIC express letters. Part B, Applications: an international journal of research and surveys* 7, 7 (2016), 1567–1574.

[22] Ksenia Shubina and John K Tsotsos. 2010. Visual search for an object in a 3D environment using a mobile robot. *Computer Vision and Image Understanding* 114, 5 (2010), 535–547.

[23] Hedvig Sidenbladh, Michael J Black, and David J Fleet. 2000. Stochastic tracking of 3D human figures using 2D image motion. In *European conference on computer vision*. 702–718.

[24] Thiago Teixeira, Gershon Dublon, and Andreas Savvides. 2010. A Survey of Human-Sensing: Methods for Detecting Presence, Count, Location, Track, and

Identity. *Comput. Surveys* 5 (01 2010), 59.
[25] Stefan Waldherr, Roseli Romero, and Sebastian Thrun. 2000. A gesture based interface for human-robot interaction. *Autonomous Robots* 9, 2 (2000), 151–173.
[26] Kaiyang Zhou, Adeline Paiement, and Majid Mirmehdi. 2017. Detecting humans in RGB-D data with CNNs. In *Machine Vision Applications (MVA), 2017 Fifteenth IAPR International Conference on*. IEEE, 306–309.