Reasoning on Shared Visual Perspective to Improve Route Directions

Jules Waldhart, Aurélie Clodic, Rachid Alami

Reasoning on Shared Visual Perspective to Improve Route Directions

Jules Waldhart[1], Aurélie Clodic[1] and Rachid Alami[1]

***Abstract*— We claim that the activity consisting in providing route directions can be best dealt with as a joint task involving the contribution not only of the robot as a direction provider but also of the human as listener. Moreover, we claim that in some cases, both the robot and the human should move to reach a different perspective of the environment which allows the explanations to be more efficient. As a first step toward implementing such a system, we propose the SVP (Shared Visual Perspective) planner which searches for the right placements both for the robot and the human to enable the visual perspective sharing needed for providing route direction and which makes the choice of the best landmark when several are available. The shared perspective is chosen taking into account not only the visibility of the landmarks, but the whole guiding task.**

I. Introduction

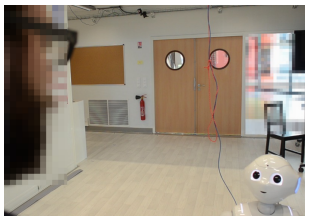
When one asks a direction to an employee in charge of providing information to visitors of a public place, said employee will most likely point a direction and give some instructions to reach your destination ("This way, take the first street on your left,..."). In trivial cases, she/he will point directly at the destination ("It is just here"). In some other interesting cases however, the employee may move and take you to a position where she/he can show you some (previously hidden) landmark ("It is just behind this corner"), thus simplifying the directions and easing your task. This is the case for our robot in the example shown in Fig. 1. These scenarios of a robotic guide could be summarized as follows:

- an interactive robot, placed near an information desk in a public space, is available to provide information and route directions
- it can move a little (say several meters around its base) in order to place itself and ask its human addressee to move with it in order for both of them to reach a configuration where it can point to one (or several) landmark(s) and utter route direction information
- the robot is not intended to accompany the persons to their destination but to help way-finding.

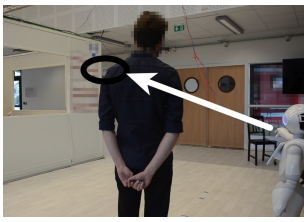
This scenario is similar to the one proposed in [1], but with a major difference: the robot is able to compute a placement for both participants, the human and the robot which offers a perspective that is more pertinent to provide route direction anchored on visible landmarks.

We will show the pertinence of choosing and reaching a different shared perspective for the route explanation by means of a few – yet significant – examples. In those examples, the robot autonomously computed a location, went there with the human and provided route directions supported by deictic gestures. We will present the core decisional component, the Shared Visual Perspective (SVP) Planner, that computes the locations (2D positions of both human and robot) to reach. The SVP Planner has been integrated in a larger system that allows the effective achievement of the full guiding task on a real robot, but this is not in the scope of this paper, neither are the complete task design details; we focus here on the SVP planner and its pertinence.

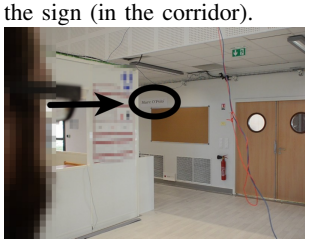
(a) Initial situation, the visitor asks for a shop.



(b) Visitor's perspective at his initial position: he cannot see the sign (in the corridor).



(c) Robot asked the human to move a little and also placed itself such as the sign is visible to both and it can point at it.



(d) Visitor's perspective from the planned position. He can see the sign now.

Fig. 1: In-lab demonstration. The robot has to show the circled landmark to the human; the SVP planner has found nearby positions for the human and the robot (c) from where it can be seen and pointed at.

Our approach differs from previous contributions and systems concerning robot guides. From the first museum guides [2], [3], [4] to more recent robot guides in large areas [5], [6], [7], the focus was more to open the road or accompany a person or a group until they reach a final destination. Here the problem is different, the robot is not authorized to move too far from its base and its role is to provide route information using gesture and speech.

While a number of issues have been studied and proposed to build and evaluate direction-giving robot behaviors, very little has been done when the robot and the human are placed

*The research leading to these results has received funding from the European Unions H2020 programme under grant agreement No. 688147, MuMMER http://mummer-project.eu/

[1]Authors are with LAAS-CNRS, Université de Toulouse, CNRS, Toulouse, France firstname.name at laas.fr

in a way where they cannot see the landmark. Indeed most of the existing work assume that they are already placed in a favorable position and, if the human is not correctly placed, they assume that she/he will adjust.

In this task, not only the robot action needs to be taken into account but also an action to be achieved by the human since they will create a mental model of the route, interpret the information, search for it in the environment, etc...[8]. This is why, we can consider that it is typically a human-robot joint task [9], [10], [11] where the robot needs to have the abilities to estimate the perspective of the human, and to elaborate a shared plan involving the human and the robot that will allow to place both of them in a desired perspective.

We focus here on the selection of a shared visual perspective for providing route directions, but this work is part of an overall project that aims to implement and evaluate a complete system for the guiding task. This involves the development of a number of other components such as a human perception system [12], a human-aware reactive motion planner [13], a Human-Robot joint action supervision [14] and the associated dialogue [15].

II. Related work

We review here some contributions related to the tasks of guiding, providing route directions and pointing. Both human cognition studies and robotic or system implementations are briefly discussed.

Landmarks selection

Landmarks are used to support route description, and it is not enough to use them, one must choose them accordingly [16], [17], [18]. The criteria for choosing landmarks are related to semantic properties, perception salience and the appeal to context [17]. Also studies show the importance and relevance of propositions connecting landmarks and the actions to take (like "at the parking lot, turn right") [18]. [8] proposes "best practices" for the choice of route direction based first on a temporospatial ordering of the statements and then on the use of shared knowledge to convey common ground during the interaction.

Pointing

In situated dialog, physical signals intended to direct the addressee attention to an element of the environment can be sorted into two classes: "directing-to" and "placing-for" [19]. In [20] a study is conducted to highlight the rich design space for deictic gestures and the necessity to adapt them to physical, environmental, and task contexts. Another key aspect for the synthesis of the robot pointing gestures is their legibility [21]. Interesting studies have been done in the analysis of gestures accompanying verbal route directions [22].

Pointing can also be seen as a joint-action where the guide has to verify that the visitor has successfully looked at the pointed direction or object, through gaze analysis and dialogue.

Placements to share visual perspective

Beyond extending one's arm, pointing at an object may require repositioning the agents to facilitate the perspective sharing and communication between the visitor and the guide [23], [24], [25].

[26] provides a pertinent analysis of the stages to a successful pointing gesture. They mention the need for the viewer to be able to see both the gesture and the referent as well as the necessity of holding the gesture until coming to mutual agreement with the observer about what is being pointed at.

The consideration of the point of view of the observer by the speaker is discussed in [23] and in [24]. In [25] the importance and role of the "Shared visual space" is stressed.

Concerning issues linked to placement planning, there are substantial results on planning sensor placement (e.g. [27]) as well as planning the robot position to let it share the human visual perspective [28] but we have found no contribution on planning shared perspective for both the human *and* the robot i.e. searching for a reachable placement of both partners.

A preliminary study [29] focuses on the way a guide and a visitor place themselves and possibly move during the explanation of a route in the context of a large mall.

Route direction

In this activity, the guide gives indications on how to reach the desired destination, mostly through dialogue, but this can be improved by some gestures. Once the route directions have been successfully communicated to the visitors, they can navigate to their destination.

The synthesis of a combination of speech and gesture in order to achieve deictic reference has been discussed in [30]. [31] proposes a model for a robot that generates route directions by integrating three crucial elements: utterances, gestures, and timing.

III. The SVP planner

The problem addressed here is related to several modalities: speech, gestures (including deictic) and navigation. These modalities are deeply connected, in the sense that they support each other: speech describes a navigation path; gestures improve speech by anchoring it to landmarks or describing actions; navigating closer to the destination simplifies the speech and may allow different (better) deictic gestures. Altogether, route directions are improved by pointing at pertinent landmarks. The guide can take the visitors to a location where said landmarks are "sufficiently" visible from their (shared) perspective.

All in all, there is a continuity of solutions between providing route directions from the starting point to guiding the visitor to his destination, including guiding only to a good perspective where to provide route directions.

The task as we address it is the sequence:

1) guiding –physically accompanying– the visitor to some place;
2) pointing at a landmark;

3) providing route directions –based on the pointed landmark;
4) reaching the destination (visitor only);

in that order, but with each step being optional. The SVP planner solves the problem of finding a position for visitor and the guide where a pointing of some landmark(s) can be performed. The landmarks to point at is dependent on the task. It is important to notice that all of these steps are taken into account by the SVP planner to evaluate the task solution as a whole[1].

A. Model

Our approach relies on a variety of information about the environment and the agents, either symbolic, physical, or on mental states. Provided with these data, the SVP planner can be potentially adapted to any situation where a robot has to provide route directions and to point at landmarks, like streets, museums, malls, offices, university campuses...

1) Physical Environment Model: The environment needs to be represented in three dimensions, its accuracy influences the pertinence of the visibility computations and navigation planning. All the obstacles to navigation or sight (occlusion) must be represented. The model must discriminate landmarks from each other and from other objects or obstacles to allow the computation of a specific landmark visibility. In our implementation, we represent the environment (including landmarks) using 3D meshes, visibility of objects is computed with OpenGL (similar to what is used by [28]).

2) Symbolic Environment Model: The SVP planner needs information at symbolic level, mostly about landmarks. The SVP planner takes as input a list of landmarks that could suit the destination. Each landmark is associated to a scalar representing the duration of the utterance of the route direction based on this landmark. [32] presents an environment model built for providing route directions that can be suitable to our approach. A similar system is being developed within our team that computes route directions and provides the pertinent landmarks to use and their evaluation.

3) Human Model (visitor): We want the guide to adapt to different human visitor capabilities, so the system is accessible and does not discriminate certain persons by ignoring their specificities, and also adapt to a range of use cases. Our system can make use of the following information to adapt the solutions:

- height of the subject eyes, to compute its perspective accordingly;
- visual acuity to enforce the use of more visible and salient landmarks;
- navigation speed, to compute plan duration and give more important penalties to long routes;
- urgency to reach the place (to balance the importance of plan duration over other criteria).

These attributes are taken as input here but we believe they can be acquired and/or inferred through dialogue and perception (e.g. persons with a stroller or loaded shopping cart, persons in a wheelchair or with crutches are usually slower than average; a person in a hurry may express it verbally or through body attitude), or updated on failure recovery[2].

4) Robot Model (guide): The robotic guide may be able to navigate, in which case the planner can take as input a maximal distance the robot can run from its initial position. We use a speed estimation to compute plan duration. Our approach can indirectly take into consideration capacities of the robot by tuning some related parameters: accuracy of the pointing gesture and gaze estimation, dialogue capacities (inducing a higher cost of dialogue-based tasks).

5) Domain Parameters: Some parameters may depend on the given domain where the robot is deployed. The guide may be allotted a limited amount of time to serve each visitor, or on the contrary be expected to help each visitor as much as possible, *i.e.* provide the maximum effort to solve a request once it has been asked to the guide.

B. Evaluating the Solutions

The decision is based on estimation and comparison of the possible solutions to the task. The solution evaluation has to take into account:

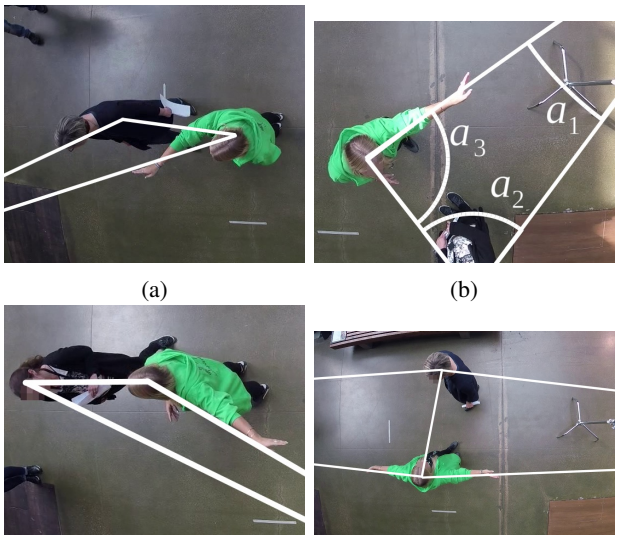
- chances of success (the simpler the indications the higher is the probability that the human will remember them and reach the destination);
- visitor effort and task duration;
- domain objectives – serving as much visitors as possible vs. providing the best quality of service for the served individuals.

1) Placements to share visual perspective: When pointing at an object, the guide objective is that the visitor identifies it unambiguously. To achieve this, it may be helpful to (1) reduce the difference of perspective, by getting the two agents almost aligned with the object. Stress is put on the alignment when the object is difficult to distinguish because it is small in the field of view. A secondary objective for the guide is to (2) relieve the visitor from some physical or mental effort by placing itself between the visitors and the destination, so they don't have to turn their head around to successively look at the pointing arm or gaze and in the pointed direction. A last objective on the pointing position is (3) for the guide to be able to monitor the visitor gaze, and speak to them; but the guide also needs to enforce pointing with gaze, so it should be able to look at the visitor and at the chosen landmark.

These properties are estimated by building a triangle whose vertices are the visitor, guide and pointed object centers, as represented in Fig. 2. The three angles (see Fig. 2b) denote the above mentioned properties of the pointing position. Angle (a_1) at the pointed object vertex correspond to the difference between the perspectives of the agents. Angle (a_2) at the visitor vertex reflects how much

[1]individual steps are heuristically evaluated on some parameters we found pertinent (mostly time) , so that their precise design should not interfere with the planner.

[2]The human could say "I don't see it", we would then replan with reduced acuity



(a) (b) (c) (d)

Fig. 2: Four examples of real pointing scenarios. The person wearing a green (light) sweatshirt is the guide pointing at a landmark (two in (d)); in white are represented the triangles formed by the landmark, visitor and guide.

they have to move to switch from looking at the guide being pointing and the pointed object; it also indicates if the guide sees the visitor's face when they look at the landmark, allowing gaze detection or not. The third angle (a_3) is for the guide to look at the object and the visitor. In the presented results, the SVP planner is configured to get $a_2 \approx a_3$ and minimize a_1.

2) Guiding: The joint navigation step is evaluated considering the distance run while guiding, and the duration of the guiding step thanks to the speed estimations provided as input. When a guiding step is necessary, the solution evaluation is penalized by a constant value that represents the time needed to ask the visitor to move and explain what she/he should do.

3) Route Directions: The utterance of route directions to the visitor, or more ambitiously the construction of a dialogue in which the directions are given to the visitor, is likely to be a time consuming step of the task. Even more importantly, it is a critical part for the success of the task: too complex instructions will be likely to lead the visitor to get lost or simply abandon the task and find another way for reaching her/his objective, making the guide counterproductive. The duration and complexity of the route directions is directly related to the number of steps of the route [18]. The guide will need to find simpler routes, use visible landmarks to simplify them, move to a place where such landmark is visible. This is illustrated in the example of Fig. 6b where the guide uses a landmark next to the door and starts its route directions from that point, hence removing one step in the route to explain (the one to reach the door from the current position). The planner will seek to choose a landmark associated to simplest possible route description

C. A Planning Request

Equipped with the data provided by the models defined section III-A, a request to SVP planner contains at least:

- initial position of the robot (guide) p_{g_0} and the human (visitor) p_{v_0},
- visitor's destination position p_{v_d},
- list of landmarks L and duration of the indication utterance based on each landmark ($T_{Indic}(l), \forall l \in L$);

and outputs:

- placements and orientation for both guide and visitor,
- list of visible landmarks from there.

Some other parameters that are set by default can be parameterized: height of the eyes of the human; height of the *eyes* of the robot: those that the human consider to be the eyes of the humanoid robot, not the camera actually used for perception; speed estimations for each agent; maximal distance the robot can run from its initial position; minimal visibility score to consider a landmark visible; other parameters to tune the importance of each aspect of the task with respect to each other in the choice of the best solution (like optimizing the duration over the visibility,...).

IV. Implementation

The SVP planner decides where the robot and human should go to reach a good shared perspective from which efficient route directions can be given. This step is a key decision of the overall task, as the position will determine all the other steps. This is why our planner uses an objective function that encompasses all the task, rather than just evaluate the quality of the pointing and perspective. Our solution is designed to match the high coupling of all the steps of the task.

The evaluation criteria presented above are represented as cost and constraints. Constraints are inequalities that represent the validity of a solution, and costs are used to choose the best solution among the valid ones.

A. Search Space

The planner decomposes the area accessible to the guide in a two-dimensional grid, and searches for the best pair of positions

$$X = (p_g, p_v) = ((x_g, y_g), (x_v, y_v))$$

both for the guide and visitor, expanding from the initial positions $X_0 = (p_{g_0}, p_{v_0})$. The destination state is $X_d = (p_{g_0}, p_{v_d})$ (the guide goes back to its initial position and the visitor reaches the destination).

B. Constraints

Constraints are computed for a tuple of : landmark, visitor position and guide position, that is (l, p_v, p_g) or equivalently (l, X). A solution is valid only if all the constraints are respected.

a) Visibility constraint: ensures that the two agents see the landmark (hence that shared perspective and joint attention are possible)

$$v(l, p_g) \geq V_{ming} \text{ and } v(l, p_v) \geq V_{minv}$$

b) Interaction distance constraint: interaction distance within 20% of the desired distance

$$(||p_v\vec{p}_g|| - D_I)^2 < (D_I \cdot 0.2)^2$$

where $||p_v\vec{p}_g||$ = distance between the agents and D_I = desired distance interaction. The value of D_I relates to the proxemics theory and is intended to ensure a social interaction distance. (The 20% error is actually a parameter.)

Guide range constraint: keeps the guide within a certain distance from its initial position

$$d_g(X) < D_{max}$$

Guide time constraint: limit the duration of the task for the guide

$$T_{Guide}(X) + T_{Return}(X) + T_{Indic}(l, X) < T_{max}$$

C. Costs

In addition to these constraints, our implementation takes the following parameters into account:

1) Navigation Distance and Duration: To estimate the distances and duration of the navigation phase, we use a the same grid as the visibility grid. It allows to compute shortest paths with Dijkstra Algorithm.We compute the distances from three points, giving distances between these points and any point in the grid. We compute distances from p_{g_0}, p_{v_0} and p_{v_d}, respectively providing the following path lengths for any X in the grid: distance runs by the guide $d_g(X) = d(p_{g_0}, p_g)$; distance run by the visitor $d_v(X) = d(p_{v_0}, p_v)$; remaining distance to reach the destination for the human $d_{destination}(X) = d(p_v(X), p_{v_d}(X))$.

We compute an estimation of the joint navigation (guiding) step duration as

$$T_{Guide}(X) = max(d_g(X)/s_g, d_v(X)/s_v)$$

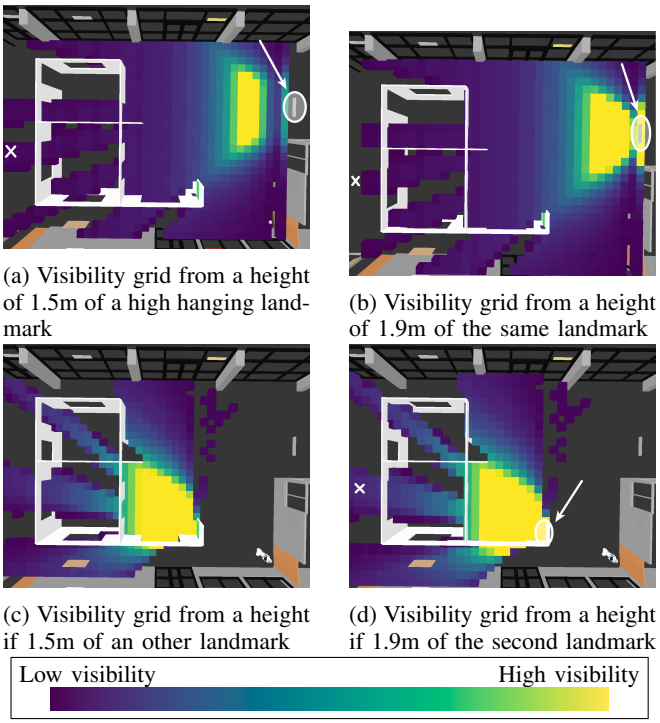
where s_g and s_v are the respective average speed estimations of the agents and the durations

$$T_{Destination}(X) = d_{Destination}(X)/s_v$$
$$T_{Return}(X) = d_s(X)/s_g$$

respectively for the human to reach the destination and for the robot to return to its base.

2) Landmarks visibility from visitor and guide placements: For each landmark l and position X we compute the visibilities of l by the guide and the visitor $v(l, p_g)$, $v(l, p_v)$.

To speed up the computation, each visibility score is precomputed, because it is a quite expensive step. The 3D space is sampled with a grid that holds score representing perceived size of the objects in the 360 degrees fields of view from each cell center (the values of $v(l, p)$ for various sizes of human). Sample visibility grid (3D) are shown in Fig. 3. The visibility computation itself is done by assigning each object a unique color, rendering the 3D scene with OpenGL and counting the number of pixel of each color, from each cell of the grid. To avoid issues related to distortion, the field of view is split in section of a maximum range of 90 degrees in each direction.



(a) Visibility grid from a height of 1.5m of a high hanging landmark

(b) Visibility grid from a height of 1.9m of the same landmark

(c) Visibility grid from a height if 1.5m of an other landmark

(d) Visibility grid from a height if 1.9m of the second landmark

Fig. 3: Grids representing the visibility of two landmarks in our lab building environment, at two different viewpoints height. The grids are actually three dimensional, we represent here two 2D slices of two visibility grids. Cells are cubes of 40cm sides. Lighter/yellow cells are those from where the visibility of the object is the best, while from dark/purple ones the object is hardly visible. Transparent cells correspond to the object being not visible at all. We see how the visibility measure is determined by distance and obstacles. The landmark for which the visibility is shown is highlighted by a white ellipse and arrow, the cross on the far left is the position from where the perspectives of Fig. 4 are taken.

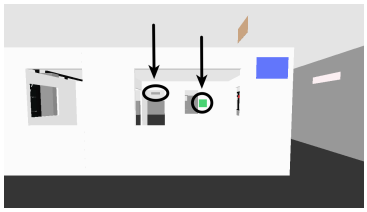
3) Route direction duration regarding a landmark: For each landmark l, providing the route direction based on that landmark has a duration estimation $T_{Indic}(l)$.

4) Pointing Conformation: We use the three angles $a_i(l, X), i = [1, 2, 3]$ representing the pointing conformation, computed from a triangle whose vertices are robot and human eyes and the center of the landmark (see Fig. 2).

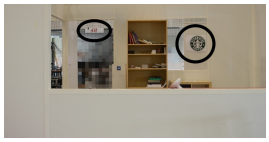
*5) **Cost Function**:* The cost function combining the parameters presented above is:

$$c(l, X) = \Big((T_{Guide}(X) + T_{Indic}(l))(K_H + K_R) + T_{Destination}(X).K_H + T_{Return}.K_R + K_v.V(l, X) + 1\Big) \times \Big(\sum_{i=1}^{3} a_i(l, X)K_{ai}\Big) \quad (1)$$

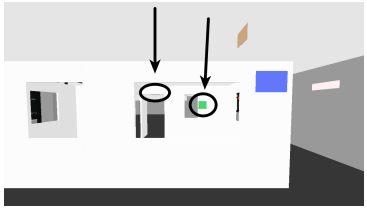
Where $V(l, X) = max(0, V_{min} - v(l, X))$, K_H is the weight applied to human time, K_R for robot, K_v is the weight applied to the visibility score, and the K_{ai} are the weights applied to each angle of the conformation.



(a) Perspective from a height of 1.5m, the two highlighted landmarks are visible.



(b) Picture of the same perspective from 1.5m height (zoomed).



(c) Perspective from a height of 1.9m, the hanging landmark is hidden by the part of the wall above the door.

(d) Picture of the perspective from 1.9m height.

Fig. 4: Two perspective taken from the position marked in Figure 3, with the same landmarks indicated by a black ellipse and arrow.

This is the cost for a landmark and position. As we want to choose the best landmark to point at, the cost $c(X)$ at a position X is the best of the $c(l, X)$, that is

$$c(X) = \min_{l \in L}(c(l, X))$$

where L is the set of landmarks provided in the request.

D. Search Algorithm

Our implementation performs a search by propagation from the cell containing X_0. The propagation is based on a set of *open* cells, where neighbors of previously closed cell are added, except when the closed cell break some evaluation constraints (namely, the guide range and time constraints). This prevents the algorithm to explore all the possibilities.

E. Choose the Best Route

One step further, the planner could be provided multiple alternative routes, and choose the best one based on the already existing cost. Indeed, we try to capture the whole task in this cost. So this would be achieved by simply running the planner for each route, and picking the one which provides the solution with the best cost.

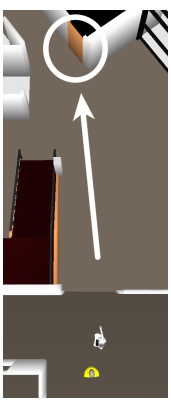
V. Examples

We present examples in two environments.

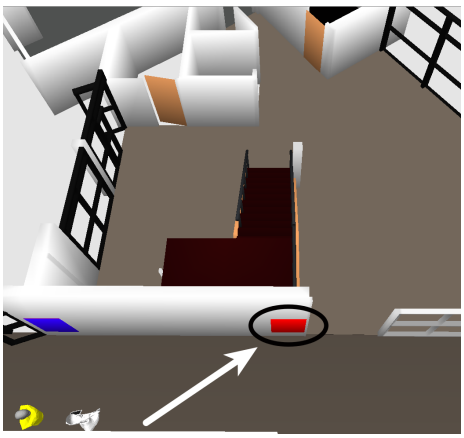
The first example (Fig. 5) is the ground floor of a building of our lab, featuring an entry hall, offices and meeting room, and a large hall with an experimenting apartment. The main hall is around 12 by 20 meters the central apartment occupies a 9x9 meters square.

In Fig. 6 we show how the robot can make use of landmarks situated on the path to the destination and balance

Fig. 5: Overview of our lab building 3D model, with virtual signs added to serve as landmarks.



(a) The robot indicates the restroom door (top of the picture) to the human



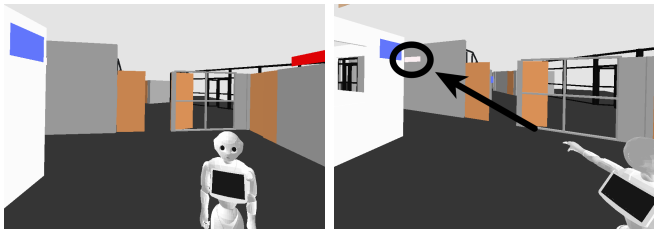
(b) The robot points at a sign to indicate where the human should go to approach and see the restroom door (both come from the left side of the picture).

Fig. 6

between guiding and providing route directions. In Fig. 6a the robot is guiding the visitor to a place where the destination is visible, leading to simple route directions; whereas in Fig. 6b, we set a low speed to the robot, so the planners prefers not to guide the human, and use a landmark to indicate a waypoint for human navigation (while the robot is still able to guide the human, the planner prefers no to).

The SVP planner has been integrated to a system allowing its execution on a real robot, we have been running in-lab demonstrations, and we plan to bring it to real life situations for testing. Fig. 1 shows pictures taken during an in-lab test, and Fig. 7 shows the human initial and planned perspectives in the 3D model. The request in this case was made of only one landmark, the shop front, and the robot asks the human to move a bit to reach a perspective where he can see the destination landmark. The robot too is moving to have a similar perspective. While the robot do not need to see the object, it is important that both agents share a visual perspective for an effective pointing.

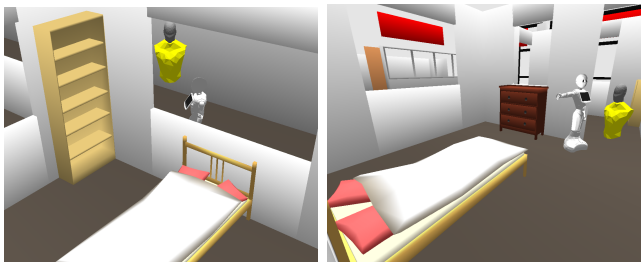
In this same environment, Fig. 8 illustrates the ability to take into account different human morphologies and adapt



(a) Initial human perspective in our 3D model.

(b) Planned perspective in the 3D model.

Fig. 7: Perspectives in the 3D model for the example of Fig. 1



(a) In this solution, the robot shows the bed to the human through a "window", limiting the joint navigation length.

(b) With the same initial situation but with a small person (child) who cannot see the bed through the window, the robotic guide guides to get in the bed room.

Fig. 8: Two distinct solutions to the same problem caused only by a different morphology of the visitor.

to their perspective when pointing at an object that can be hidden by obstacles, leading to very different solutions, in this case with a small child unable to look over a window edge.

We have been testing the system in a larger environment, a mall, where the robot can navigate only in a small area of a vast hall (Fig. 9 and 10). The solution is to guide the human to a place from where he can see a landmark close to the target (as close as possible). If the robot wouldn't move, it would have to point at a landmark further away from the target (say, here, the green *info* panel, visible near the middle of Fig. 9 and in Fig. 10).

In these test, when the supervising component detects the human did not moved at the planned position, it requests the SVP Planner an evaluation of this position (applying the cost function to it). If it is valid, then the execution continues with that position, otherwise the guiding starts again from that position[3].

VI. Conclusions

We have shown in this paper that we can compute a place where a robot can accompany a human to provide route directions based on landmarks that would otherwise be invisible. Doing so may be pertinent when such landmark

[3]Details of this implementation are out of the scope of this publication, but it appeared to be an important concern in most reviews.

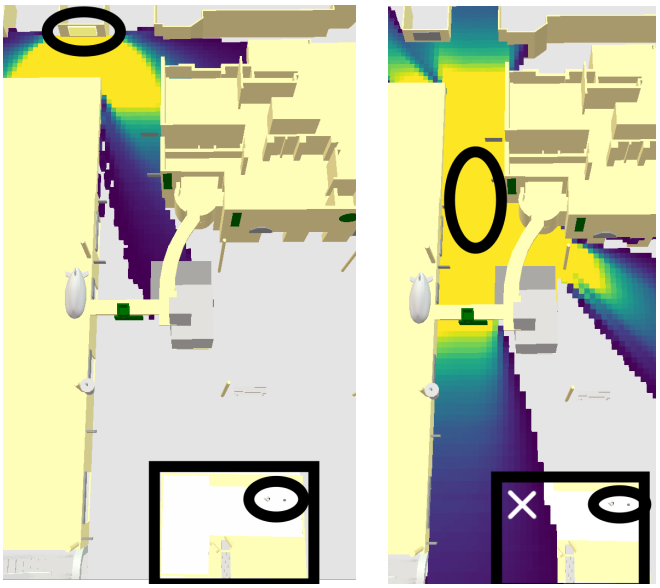


Fig. 9: Bird's-eye view of the real-life mall visibility grids of the target landmark (left) and a waypoint landmark (right). The bottom black box is the area allowed for the robot to navigate, the circle in it highlights the starting position of the experiment in Fig. 10, where the robot guides the human near the position highlighted by the white cross mark.

Fig. 10: Real-life mall experiment video screen-shots: (left) initial position; (right) robot pointing at the corridor, after guiding the human a few meters to reach an acceptable shared visual perspective.

may help improve the quality of the indications, as shown by some examples. This approach requires to consider the problem as a joint task, as we produce a shared plan where both agents need to act.

The preliminary implementation of a solver dedicated to this task, the SVP Planner, gives us a view of the requirements for implementing a system aiming at tackling this complete task, from request to execution. Our planner needs information about the route to indicate to the visitor: the path(s) it can take, and landmarks that could improve the route directions if they can be pointed at. Knowledge about the visitor goal, mental state and capacities presented in III-A.3 can be provided by dedicated tools based on dialogue and visual perception. The execution of the navigation (guiding) step is widely addressed in the literature. The pointing

gesture by itself is also addressed, along with the association of gestures with verbal route directions. These elements would work with objectives provided by the SVP planner presented in this paper: guiding destination, landmarks to point and route to indicate. Execution of the task requires dedicated supervision to articulate the various phases of the interaction and eventually recover from failures, *e.g.* by re-planing with updated parameters. We are developing such a complete system in the scope of the MuMMER project.

REFERENCES

[1] T. Kanda, M. Shiomi, Z. Miyashita, H. Ishiguro, and N. Hagita, "A Communication Robot in a Shopping Mall," *IEEE Transactions on Robotics*, vol. 26, no. 5, pp. 897–913, Oct. 2010.

[2] W. Burgard, A. B. Cremers, D. Fox, D. Hähnel, G. Lakemeyer, D. Schulz, W. Steiner, and S. Thrun, "The museum tour-guide robot RHINO," in *Autonome Mobile Systeme 1998, 14. Fachgespräch, Karlsruhe, 30. November - 1. December 1998*, pp. 245–254.

[3] A. Clodic, S. Fleury, R. Alami, R. Chatila, G. Bailly, L. Brethes, M. Cottret, P. Danès, X. Dollat, F. Elisei, I. Ferrané, M. Herrb, G. Infantes, C. Lemaire, F. Lerasle, J. Manhes, P. Marcoul, P. Menezes, and V. Montreuil, "Rackham: An interactive robot-guide," in *The 15th IEEE International Symposium on Robot and Human Interactive Communication, RO-MAN 2006, Hatfield, Herthfordshire, UK, 6-8 September, 2006*, pp. 502–509.

[4] R. Siegwart, K. O. Arras, S. Bouabdallah, D. Burnier, G. Froidevaux, X. Greppin, B. Jensen, A. Lorotte, L. Mayor, M. Meisser, R. Philippsen, R. Piguet, G. Ramel, G. Terrien, and N. Tomatis, "Robox at expo.02: A large-scale installation of personal robots," *Robotics and Autonomous Systems*, vol. 42, no. 3-4, pp. 203–222, 2003.

[5] R. Kümmerle, M. Ruhnke, B. Steder, C. Stachniss, and W. Burgard, "A navigation system for robots operating in crowded urban environments," in *IEEE International Conference on Robotics and Automation, Karlsruhe, Germany*, 2013, pp. 3225–3232.

[6] A. M. Bauer, K. Klasing, T. Xu, S. Sosnowski, G. Lidoris, Q. Mühlbauer, T. Zhang, F. Rohrmüller, D. Wollherr, K. Kühnlenz, and M. Buss, "The autonomous city explorer project," in *2009 IEEE International Conference on Robotics and Automation, ICRA 2009, Kobe, Japan, May 12-17, 2009*, pp. 1595–1596.

[7] R. Triebel, K. Arras, R. Alami, L. Beyer, S. Breuers, R. Chatila, M. Chetouani, D. Cremers, V. Evers, M. Fiore, H. Hung, O. A. I. Ramírez, M. Joosse, H. Khambhaita, T. Kucner, B. Leibe, A. J. Lilienthal, T. Linder, M. Lohse, M. Magnusson, B. Okal, L. Palmieri, U. Rafi, M. van Rooij, and L. Zhang, "SPENCER: A Socially Aware Service Robot for Passenger Guidance and Help in Busy Airports," in *Wettergreen D., Barfoot T. (Eds) Field and Service Robotics*, ser. Springer Tracts in Advanced Robotics. Springer, Cham, 2016, pp. 607–622.

[8] G. L. Allen, "Principles and practices for communicating route knowledge," *Applied cognitive psychology*, vol. 14, no. 4, pp. 333–359, 2000.

[9] N. Sebanz, H. Bekkering, and G. Knoblich, "Joint Action: Bodies and Minds Moving Together," *Trends in Cognitive Sciences*, vol. 10, no. 2, pp. 70–76, 2006.

[10] S. Lemaignan, M. Warnier, E. A. Sisbot, A. Clodic, and R. Alami, "Artificial cognition for social human–robot interaction: An implementation," *Artificial Intelligence*, vol. 247, 2017.

[11] A. Clodic, E. Pacherie, R. Alami, and R. Chatila, "Key elements for human-robot joint action," in *Sociality and Normativity for Robots*, R. Hakli and J. Seibt, Eds. Springer International Publishing, 2017, pp. 159–177, DOI: 10.1007/978-3-319-53133-5_8.

[12] S. Duffner and J.-M. Odobez, "Track creation and deletion framework for long-term online multiface tracking," *IEEE Transactions on image processing*, vol. 22, no. 1, pp. 272–285, 2013.

[14] S. Devin and R. Alami, "An implemented theory of mind to improve human-robot shared plans execution," in *The Eleventh ACM/IEEE International Conference on Human Robot Interation, HRI 2016, Christchurch, New Zealand, March 7-10, 2016*, pp. 319–326.

[13] H. Khambhaita and R. Alami, "Viewing Robot Navigation in Human Environment as a Cooperative Activity," in *International Symposium on Robotics Research (ISSR 2017)*, Puerto Varas, Chile, 2017, p. 18p.

[15] I. Papaioannou, C. Dondrup, J. Novikova, and O. Lemon, *Hybrid chat and task dialogue for more engaging HRI using reinforcement learning*. IEEE, Aug 2017, p. 593598. [Online]. Available: http://ieeexplore.ieee.org/document/8172363/

[16] C. Nothegger, S. Winter, and M. Raubal, "Selection of Salient Features for Route Directions," *Spatial Cognition & Computation*, vol. 4, no. 2, pp. 113–136, 2004.

[17] H. H. Clark, R. Schreuder, and S. Buttrick, "Common ground at the understanding of demonstrative reference," *Journal of verbal learning and verbal behavior*, vol. 22, no. 2, pp. 245–258, 1983.

[18] M.-P. Daniel, A. Tom, E. Manghi, and M. Denis, "Testing the Value of Route Directions Through Navigational Performance," *Spatial Cognition & Computation*, vol. 3, no. 4, pp. 269–289, Dec. 2003.

[19] H. H. Clark, "Coordinating with each other in a material world," *Discourse Studies*, vol. 7, no. 4, pp. 507–525, Oct. 2005.

[20] A. Sauppé and B. Mutlu, "Robot deictics: How gesture and context shape referential communication," in *Proceedings of the 2014 ACM/IEEE International Conference on Human-robot Interaction*, ser. HRI '14. New York, NY, USA: ACM, 2014, pp. 342–349.

[21] R. M. Holladay, A. D. Dragan, and S. S. Srinivasa, "Legible robot pointing," in *Robot and Human Interactive Communication, 2014 RO-MAN: The 23rd IEEE International Symposium On*. IEEE, Aug. 2014, pp. 217–223.

[22] G. L. Allen, "Gestures accompanying verbal route directions: Do they point to a new avenue for examining spatial representations?" *Spatial cognition and computation*, vol. 3, no. 4, pp. 259–268, 2003.

[23] M. W. Alibali, "Gesture in Spatial Cognition: Expressing, Communicating, and Thinking About Spatial Information," *Spatial Cognition & Computation*, vol. 5, no. 4, pp. 307–331, Dec. 2005.

[24] W.-T. Fu, L. D'Andrea, and S. Bertel, "Effects of Communication Methods on Communication Patterns and Performance in a Remote Spatial Orientation Task," *Spatial Cognition & Computation*, vol. 13, no. 2, pp. 150–180, Apr. 2013.

[25] J. W. Kelly, A. C. Beall, and J. M. Loomis, "Perception of shared visual space: Establishing common ground in real and virtual environments," *Presence: Teleoperators & Virtual Environments*, vol. 13, no. 4, pp. 442–450, 2004.

[26] N. Wong and C. Gutwin, "Where are you pointing?: The accuracy of deictic pointing in CVEs," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM Press, 2010, pp. 1029–1038.

[27] S. Y. Chen and Y. F. Li, "Automatic sensor placement for model-based robot vision," *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 34, no. 1, pp. 393–408, Feb 2004.

[28] L. F. Marin-Urias, E. A. Sisbot, A. K. Pandey, R. Tadakuma, and R. Alami, "Towards shared attention through geometric reasoning for Human Robot Interaction," in *2009 9th IEEE-RAS International Conference on Humanoid Robots*. IEEE, Dec. 2009, pp. 331–336.

[29] K. Belhassein, A. Clodic, H. Cochet, M. Niemel, P. Heikkila, H. Lammi, and A. Tammela, "Human-human guidance study," LAAS-CNRS, CLLE, VTT, tech-report hal-01719730, 12 2017.

[30] A. G. Brooks and C. Breazeal, "Working with robots and objects: Revisiting deictic reference for achieving spatial common ground," in *Proceedings of the 1st ACM SIGCHI/SIGART Conference on Human-robot Interaction*, ser. HRI '06. New York, NY, USA: ACM, 2006, pp. 297–304.

[31] Y. Okuno, T. Kanda, M. Imai, H. Ishiguro, and N. Hagita, "Providing route directions: Design of robot's utterance, gesture, and timing," in *Human-Robot Interaction (HRI), 2009 4th ACM/IEEE International Conference On*. IEEE, 2009, pp. 53–60.

[32] Y. Morales, S. Satake, T. Kanda, and N. Hagita, "Building a Model of the Environment from a Route Perspective for Human–Robot Interaction," *International Journal of Social Robotics*, vol. 7, no. 2, pp. 165–181, Apr. 2015.