IAV2004 - PREPRINTS
*5th IFAC/EURON Symposium on Intelligent Autonomous Vehicles*
*Instituto Superior Técnico, Lisboa, Portugal*
*July 5-7, 2004*

# LEARNING NAVIGATION MAPS BY LOOKING AT PEOPLE

**Roger Freitas** [*,1] **José Santos-Victor** [**]
**Mário Sarcinelli-Filho** [*] **Teodiano Bastos-Filho** [*]

[*] *Departamento de Engenharia Elétrica, Universidade Federal do Espírito Santo, Vitória, ES - Brasil*
[**] *Instituto de Sistemas e Robótica, Instituto Superior Técnico, Lisboa - Portugal*

Abstract: Mobile robots remain idle during significant amounts of time in many applications, while new tasks are not assigned. In this paper, we propose a framework to use those periods of inactivity to observe the surrounding environment and *learn* information that can be used later on during navigation. Events like someone entering or leaving a room, or someone approaching a printer to pick a document up, convey important information about the observed space and the role played by the objects therein. We explore the information implicitly present in the motion patterns people describe in a certain workspace, to allow the robot to infer a "meaningful" spatial description. Map building is thus bottom-up driven by the observation of human activity, and not simply a top-down oriented geometric construction.

Keywords: Learning, Imitation, Navigation, Topological Navigation

## 1. INTRODUCTION

In many applications, mobile robots remain idle for significant amount of time, while a new task is not assigned. Similarly, in many research labs, mobile robots remain inactive during extended periods of time, while new sensor processing or navigation algorithms are being tested.

The motivation of this work is to use those periods of inactivity to observe the surrounding environment and *learn* information that can be used later on during navigation. For example, events like someone entering or leaving a room, someone approaching a printer to pick a document up, or a group of people have a meeting around a table, convey important information about the observed space and the role played by the objects therein.

The development of algorithms to extract useful information from the observation of such events could bring significant savings in programming, while affording the robot with an extended degree of flexibility and capacity of adaptation.

In this work, we explore the information implicitly present in the motion patterns people describe in a certain workspace, to allow the robot to infer a "meaningful" spatial description. Interestingly, such spatial representation is not driven by abstract geometrical considerations but, rather, by the role or function associated to locations or objects and learnt by observing people's behaviour.

The mobile robot we use in this work, combines peripheral and foveal vision. The peripheral vision is provided by an omnidirectional camera that captures the attention stimuli to drive a standard narrow field of view pan-tilt (perspective) camera (foveal vision).

Other research groups have used information associated to people's trajectories to help robot navigation. In (Bennewitz *et al.*, 2002), mobile robots equipped with laser sensors are used to extract trajectories of people moving in houses and offices. The trajectories were estimated using

[1] Supported by *CAPES - Coordenação de Aperfeiçoamento de Pessoal de Nível Superior* (Brazil) and project POSI/SRI/41506/261 Omnisys through *Fundação para a Ciência e Tecnologia*(Portugal).

the EM algorithm and the models were used to predict human trajectories in order to improve people following. In (Bennewitz *et al.*, 2003) the same authors proposed a method for adapting the behavior of a mobile robot according to the activities of the people in its surrounding. In (Kruse and Wahl, 1998), an off-board camera-based monitoring system is proposed to help mobile robot guidance. In (Appenzeller *et al.*, 1997), it is developed a system that builds topological maps by looking at people. Their approach is based on cooperation between intelligent spaces and robots. Our approach to this problem is to extract the motion patterns of people from the robot's viewpoint directly, using an on-board camera system. The advantage of our approach is that the robot can learn from environments that are not structured for this purpose.

In Section 2 we describe the overall learning system. Preliminary results are shown in Section 3. Then, we present some conclusions and discuss possible developments.

## 2. OVERALL APPROACH

Figure 1 shows a scheme of our overall approach. The most important subsystems are the vision, measurement and modeling systems, with increasingly higher level of cognition:

- the *vision system* comprises both peripheral and foveal visual capabilities. Peripheral vision is accomplished by an omnidirectional camera and is responsible for movement detection. Foveal vision is accomplished by a perspective camera that is able to execute pan and tilt rotations. Foveal vision is responsible for tracking moving objects;
- the *measurement system* is responsible for transforming visual information into features the robot is trying to learn, e.g., transforming 2D image information into trajectory points on the floor, referred to a common coordinate frame;
- the *modeling system* is responsible for building models that explain data from the measurement system. Depending on the kind of model the robot is trying to build, the modeling system could drive the way vision system operates (e.g. controlling the gaze direction).

In this paper, we use the scheme shown in Figure 1 to learn possible trajectories and interesting places in the robot's environment. In this case, the Measurement System is responsible for transforming 2D image information into trajectory points on the floor. The Modeling System is responsible for building models of possible trajectories and/or

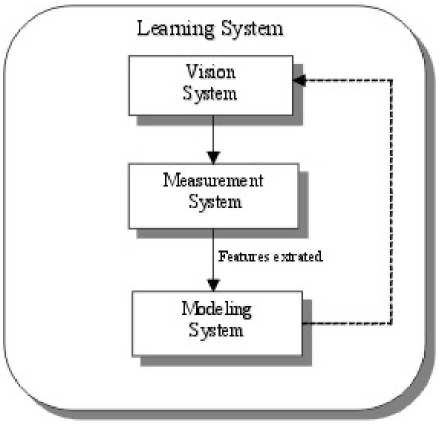


Fig. 1. The Learning System

finding interesting places in the environment that should be investigated in more detail.

We assume that the robot does not have any *prior* knowledge about the environment. From any general position, the robot goal is to extract information which allows it to navigate in the environment. In order to do that, the robot must be able to detect moving objects, track these objects and transform this information into possible trajectories (a set of positions in an external coordinate system) to be followed.

In the following sections, we describe in detail each of the subsystems we have been discussing.

### *2.1 Vision System*

The vision system comprises two types of visual information: peripheral and foveal (see Figure 2). The peripheral vision uses an omnidirectional camera, which is the primary sensor for the detecting interesting image events and to drive the attention of the foveal camera. The foveal vision system is then used for tracking the objects of interest, using a perspective camera with a pan-tilt platform.

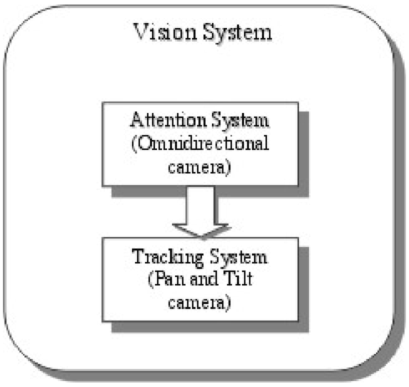

Vision System
Attention System (Omnidirectional camera)
Tracking System (Pan and Tilt camera)

Fig. 2. The Vision System

*2.1.1. Attention System* The attention system operates in the omnidirectional images and detects motion of objects or people in the robot vicinity. Other visual cues could be considered but, in the current stage of implementation, we rely exclusively on motion information.

Motion detection can be easily performed by using background subtraction. Moving objects are detected by subtracting the current image from the background image (previously obtained).

One of the several ways to obtain a representation for the background is to capture an image of the environment surrounding the robot while nobody is in the field of view of the cameras. Obviously, such method would be too restrictive, only applicable in very controlled environments.

Since in our approach the robot gathers knowledge by observing people's movements, it must be able to extract the background model even when people are moving nearby. There are several algorithms proposed in the literature that address this problem. In this work, the background is modeled using the method proposed in (Gutchess *et al.*, 2001), which uses a sequence of images taken from the same place and outputs a statistical background model describing the static parts of the scene. Multiple hypotheses of the background value at each pixel are generated by locating periods of stable intensity in the sequence. The likelihood of each hypothesis is then evaluated using optical flow information from the neighborhood around the pixel, and the most likely hypothesis is chosen to represent the background. An example of background modeling can be seen in Figure 3.

Figure 4 shows an omnidirectional image taken in the laboratory and the result of movement detection. Once the movement is detected, a command is sent to the pan and tilt camera to drive its gaze direction towards the region of interest and to start tracking the moving object. As the two cameras are fixed at the top of the robot, we can previously calibrate the transformation relating the two cameras *in terms of pan angle.* We assume that tilt angle in home position will suffice when trying to put the moving object in the field of view of the perspective camera.

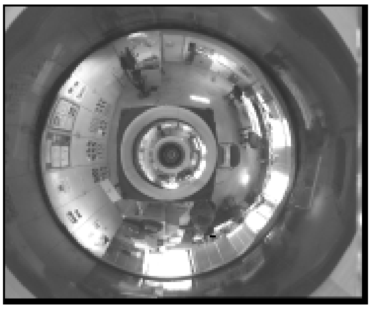
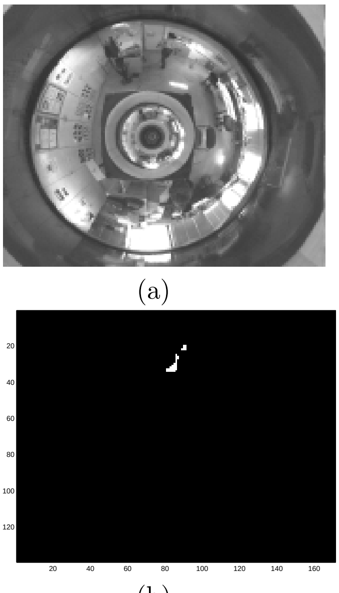


Fig. 3. Background extracted from a sequence of images taken from the laboratory.


(a)

(b)

Fig. 4. Omnidirectional image captured (a), movement detection (b).

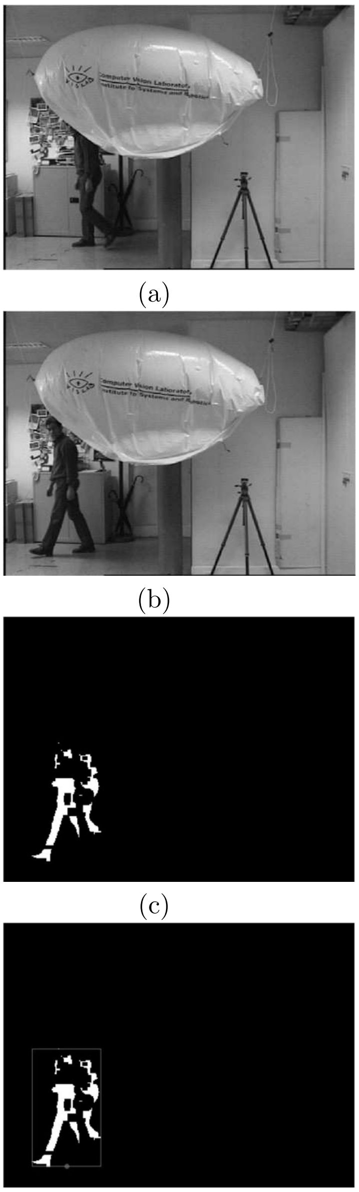
*2.1.2. Tracking System* We are currently using a simple tracking algorithm to illustrate the concept of learning about the environment from observing human actions. We are currently working to improve its performance and robustness.

The tracking routine takes two consecutive images as the input and extracts those pixels that display some change. The result is that different regions (moving objects) in the two images are highlighted (see Figure 5 (a), (b) and (c)). Then, we calculate a bounding box around the detected area. The point to be tracked is shown in gray in Figure 5 (d), i.e., the middle point of the bottom edge of the bounding box (theoretically a point on the floor).

While operating, the system is continuously detecting regions of interest in the peripheral field of view. The foveal vision system then tracks these objects, while they remain visible. The measurement system described in the following section will integrate the information of different tracked object's into a common coordinate system, where more global information can be interpreted.

### *2.2 Measurement System*

In order to estimate trajectories relative to the robot, it is necessary to estimate the distance from the robot to the moving object in each image acquired. Usually, this problem is solved using



(a)

(b)

(c)

(d)

Fig. 5. Two consecutive images (a) and (b), movement detection (c) and the point to track (d).

two or more cameras set in different places and applying stereo vision techniques.

In this work, as the robot is stationary while learning the environment, consecutive images of a given moving object differ only by camera rotations (pan and tilt). In the absence of translation, stereo cannot be used to reconstruct the 3D trajectory of the target. An alternative to solve this problem is to estimate the homography *(H)* between the floor and the image plane, i.e., to find *a priori* plane projective transformation that transforms an image point *(u,v)* into a point on the floor *(X,Y,1)* (see Equation 1).

$$\lambda \begin{pmatrix} u \\ v \\ 1 \end{pmatrix} = \mathbf{H} \begin{pmatrix} X \\ Y \\ 1 \end{pmatrix} \qquad (1)$$

where H is the $3 \times 3$ homography matrix. Initially, the homography is estimated using a set of ground plane points, whose 3D positions are known with respect to some reference frame. Then, when the foveal camera moves, the homography is updated as a function of the performed motion. So, as the camera is tracking the object, its pose is changing, and the same happens to the homography between the image plane and the floor. For this reason, we use the pan and tilt angles to update the homography (see Figure 6).

We assume that the pan-tilt camera's intrinsic parameters are known *a priori*, possibly obtained from an initial calibration step. The intrinsic parameters are used to decompose the homography matrix into a rotation matrix and a displacement vector (camera pose) relating the camera frame to a world frame. Pan and tilt angles generate canonical rotation matrices that multiply the original rotation matrix, thus updating the homography.

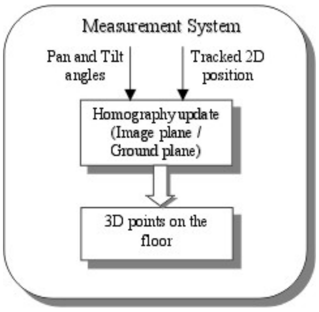


Fig. 6. The Measurement System

In order to recover camera pose, we apply the methodology presented in (Gracias and Santos-Victor, 2000), which we briefly describe next. The homography, $H$, can be written as

$$H = \lambda K L \qquad (2)$$

where $\lambda$ is an unknown scale factor, $K$ is the camera intrinsic parameter matrix and $L$ is a matrix constructed from the full (3 x 3) rotation matrix $R$ and the translation vector $t$, as follows:

$$L = \left[ \overline{R} \; t \right], \qquad (3)$$

where $\overline{R}$ is a 3x2 submatrix comprising the first two columns of matrix $R$.

Due to noise in the estimation process, homography $H$ will not follow exactly the structure of Equation 2. Alternatively, using the Frobenius norm to measure the distance between matrices, the problem can be formulated as

$$\lambda, L = \arg \underset{\lambda, L}{min} \| \lambda L - K^{-1} H \|^2_{frob} \qquad (4)$$

subject to $\overline{L}^T \overline{L} = I_2$, where $\overline{L}$ is a 3x2 submatrix comprising the first two columns of matrix $L$.

The solution of Equation 4 can be found through Singular Value Decomposition (SVD). Let $U.\Sigma.V^T$ be the SVD of $K^{-1}H$. Then, $\overline{L}$ is given by

$$\overline{L} = U.V^T \tag{5}$$

and

$$\lambda = \frac{tr(\Sigma)}{2} \tag{6}$$

The last column of $L$ can be found as

$$t = K^{-1}.H. \begin{bmatrix} 0 \\ 0 \\ \frac{1}{\lambda} \end{bmatrix} \tag{7}$$

Then

$$L = \left[ \overline{L} \; t \right] \tag{8}$$

The last column of rotation matrix $R$ can be found by computing the cross product of the the first two columns. The updated rotation matrix is given by

$$NewR = R \cdot R_{PAN} \cdot R_{TILT} \tag{9}$$

Finally, the updated homography is then

$$NewH = \lambda \cdot K \cdot NewL \tag{10}$$

where

$$\begin{aligned} NewL &= \left[ \overline{NewR} \; Newt \right] \\ Newt &= NewR \cdot t \end{aligned}$$

We now have a means to project all tracked trajectories onto a common coordinate system associated to the ground plane. In this global coordinate system, the different trajectories described by moving objects, can be further analyzed and modeled, as described in the following section.

### *2.3 Modeling System*

The modeling system is responsible for building models that explain data from the measurement system. Depending on the nature of the models the robot is building, the modeling system can drive the way vision system operates.

The modeling system aims to interpret the observed (global) trajectories onto representations that can be used for navigation. Currently, we consider two main uses of such data:

- the observed trajectories correspond to free (obstacle free) pathways that the robot may use to move around in the environment;
- if many trajectories meet in a certain area, it means that that region must correspond to some important functionality (e.g. doors, tables, tools, etc) and should be represented in a map.

Hence, from observation, the robot can learn interesting places in the scene and the most frequent ways to go from one point to another. Moving further, the robot may also be able to distinguish uncommon behaviour and thus be used in surveillance and monitoring tasks.

## 3. EXPERIMENTAL RESULTS

We performed preliminary experiments in the laboratory to verify the performance of the Vision, Measurement and Modeling systems. The robot was observing the laboratory while someone was walking along different trajectories. Each trajectory was performed and recorded separately. The positions on the floor, measured by the system, are shown in Figure 7.

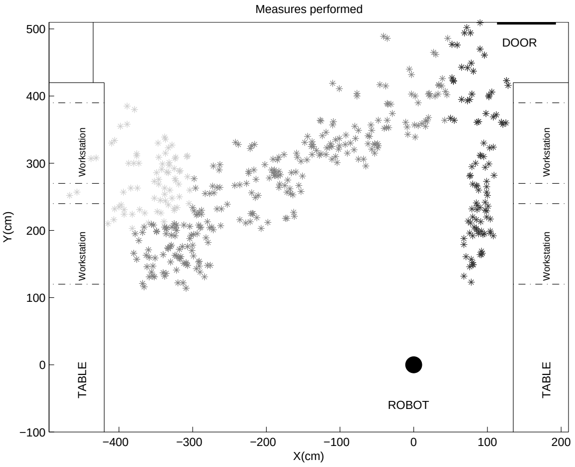


Fig. 7. Real data measured from observing people's movements.

The data generated by the Measure System is then interpreted by the Modeling System. When analyzing the data shown in Figure 7, the most interesting point is the kind of information that can be extracted from such data. For example, one can try to model this information (Bennewitz *et al.*, 2002) as possible trajectories (see Figure 8) the robot can follow whenever going from one place to another.

To illustrate the idea, the trajectories shown in Figure 8 were modeled using a linear polynomial model. In (Bennewitz *et al.*, 2002), the authors modeled the trajectories using a set of gaussian distributions.

Places of interest can be detected as well (see Figure 9. In this case, we applied a threshold on the data shown in Figure 7 based on the number of times a position was visited. This is done in order to filter the data, thus discarding positions that are not frequently visited. Then, we use k-means algorithm to cluster remaining data.

By identifying these places, a strategy for modeling and identification can be derived, thus providing an autonomous way of learning models for such places. For example, as we can see in Figure 9, two of these places of interest appear in front of workstations in the laboratory.

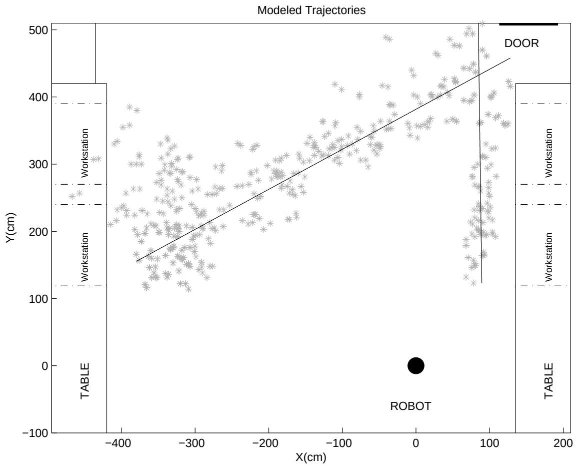


Fig. 8. Examples of possible trajectories.

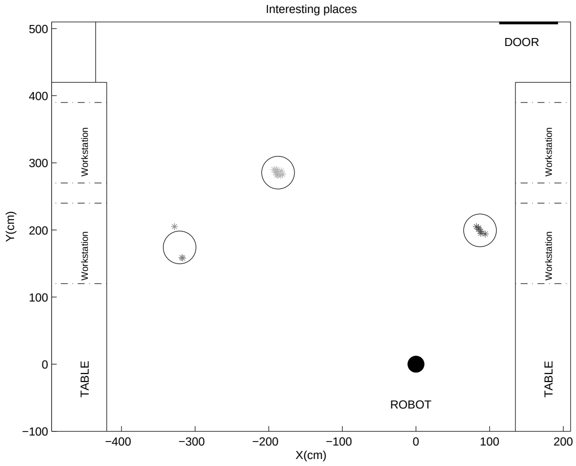


Fig. 9. Examples of places of interest.

## 4. CONCLUSIONS AND FUTURE WORK

In this paper we discussed the idea of learning how to navigate in an environment, through the observation of people's motion patterns in the same environment.

We presented an on-board vision-based approach to extract information from the environment that is useful for navigation purposes. It combines peripheral vision for attention detection and foveal vision for tracking. This information is integrated into a common coordinate system by the Measurement System. Then, the Modeling System provides higher level interpretation of the observation that can be used for navigation. Such interpretations can be in the form of feasible paths or points of interest in the scene, and be integrated in a map (e.g. a topological map).

We have presented preliminary but encouraging experimental results. We are currently focusing on improving the robustness of the system.

Throughout the paper, we consider that the robot remains static. Developing a strategy for integrating all this information (possible trajectories and places of interest), as the robot moves along, is the next step in this work. The final goal is to generate, in an autonomous and incremental way, a map that encloses useful information, allowing the robot to perform different tasks, and learn in an open-ended way, how to represent the world based on functionalities (affordances) acquired through observation.

## REFERENCES

Appenzeller, G., J. Lee and H. Hashimoto (1997). Building topological maps by looking at people: An example of cooperation between intelligent spaces and robots. *Proceedings of the International Conference on Intelligent Robots and Systems (IROS)* **3**, 1326–1333.

Bennewitz, M., W. Burgard and S. Thrun (2002). Using EM to learn motion behaviors of persons with mobile robots. *Proceedings of the International Conference on Intelligent Robots and Systems (IROS)* **1**, 502–507.

Bennewitz, M., W. Burgard and S. Thrun (2003). Adapting navigation strategies using motions patterns of people. *Proceedings of the International Conference on Robotics and Automation (ICRA)* **2**, 2000–2005.

Gracias, N. and J. Santos-Victor (2000). Underwater video mosaics as visual navigation maps. *VisLab-TR 07/2000 - Computer Vision and Image Understanding* **79(1)**, 66–91.

Gutchess, D., M. Trajković, E. Cohen-Solal, D. Lyons and A. K. Jain (2001). A background model initialization algorithm for video surveillance. *International Conference on Computer Vision* **1**, 733–740.

Kruse, E. and F. Wahl (1998). Camera-based monitoring system for mobile robot guidance. *Proceedings of the International Conference on Intelligent Robots and Systems (IROS)* **2**, 1248–1253.