IMPLEMENTATION OF HUMAN PERCEPTION ALGORITHMS ON A MOBILE ROBOT

Mathias Fontmarty [*] **Thierry Germa** [*]
Brice Burger[†] [*] **Luis Felipe Marin** [*]
Steffen Knoop [**]

[*] *LAAS - CNRS, 7 avenue du Colonel Roche - 31077*
Toulouse Cedex 4, FRANCE
[**] *Universität Karlsruhe (TH), Kaiserstraße 12 - 76131*
Karlsruhe, GERMANY
[†] *IRIT - UPS, 118 Route de Narbonne, F - 31062*
Toulouse Cedex 9, FRANCE

Abstract: During last years, a lot of works in robotic research have explored Human-Robot interactions. Hence, a great challenge in next future will be the personal robot, with perception faculties which will enable a wide range of activities such as human localization and tracking, gesture recognition and interpretation, or object manipulation.
In this paper, we will focus on human perception and we will present a human aware system implemented on a mobile robot. This system uses data from various sensors to be able to localize and to track a human presence in a wide range of distances. An exploitation of all these modalities is presented in a demo showing the robot giving an object to a person.

Keywords: Mobile Robotics, Integration, Human Perception

1. INTRODUCTION

Autonomous robots have been improved for the last years, being able to adapt to always more complex environments. In the perspective to elaborate a “personal robot”, one of the most challenging of these environments is the human one. Robots will have to be able to adapt to social environments and the human presence must be taken in account. Moreover, if we want the robot to interact with humans, and not only to move itself among them, perception abilities must be even more accurate in order to interpret visual or spoken orders, or to accomplish tasks such as exchanging an object with a person.

Thus, the work presented here mainly focuses on the perception abilities of the Jido robot, a mobile platform designed for human robot interaction, on which we work at LAAS-CNRS.

The work described in this paper is conducted within the EU Integrated Project COGNIRON (“The Cognitive Robot Companion”) and funded by European Commission Division FP6-IST Future and Emerging Technologies under Contract FP6-002020.

In section 2, we present a few technical characteristics of the robot. Section 3 details the 4 modules involved in the scenario. Results are shown and commented in section 4 and, finally, section

5 concludes our work and presents some future experiments.

2. JIDO PRESENTATION

2.1 Hardware

Jido, a MP-L655 platform from Neobotix, is a mobile robot designed to interact with human beings. It is presented on figure 1(a).

Jido is equipped with: (i) a 6-DOF arm, (ii) a pan-tilt unit system at the top of a mast (dedicated to human-robot interaction mechanisms), (iii) a 3D swissranger camera and (iv) a stereo camera, both embedded on the pan tilt unit, (v) a second video system fixed on the arm wrist for object grasping, (vi) two laser scanners, (vii) one panelPC with tactile screen for interaction purpose, and (viii) one screen to provide feedback to the robot user. Jido has been endowed with functions enabling to act as robot companion and especially to exchange objects with human beings. So, it embeds robust and efficient basic navigation and object recognition abilities.

In the human perception demonstration we present in this paper, we use the stereo camera and the swissranger device mounted on the pan-tilt unit, as well as the front laser sensor.

2.2 Software

Jido is fitted with the "LAAS" software architecture thoroughly presented in (Alami *et al.*, 1998). On the top of the hardware (sensors and effectors), the *functional level* listed in figure 1(b), encapsulates all the robot's action and perception capabilities into controllable communicating modules, operating at very strong temporal constraints. The *executive level* activates these modules, controls the embedded functions, and coordinates the services depending on the task high-level requirements. Finally, the upper *decision level* copes with task planning and supervision, while remaining reactive to events from the execution control level.

3. THE HUMAN PERCEPTION MODULES

Human perception abilities are split in 4 modules:

3.1 The Gest module

The goal of this module is to provide a 3D human hand tracking from the video stream of a stereoscopic system. Actually, the hand is modeled by a 3D deformable ellipsoid. We fit the template through the estimation of its space coordinates (x, y, z), its size (ax, ay, az), and its orientation (θ, ϕ, ψ). All these parameters are included in the state vector $x_k = (x_k, y_k, z_k, ax_k, ay_k, az_k, \theta_k, \phi_k, \psi_k)$ related to the $k - th$ frame.

As the robot's evolution takes place into dynamic and cluttered environments, several hypotheses must be handled at each instant concerning the tracker parameters to be estimated. Particle filtering thus seems well suited to this context. Moreover, in our context, the tracked hand can temporarily leave the camera field of view; that's why it needs automatic reinitialization. Therefore we based our tracker on the I-Condensation algorithm (Isard and Blake, 1998).

With regard to the dynamics model $p(x_k|x_{k-1})$, the hand motion is difficult to characterize over time. We assume that the state vector entries evolve according to mutually independent random walk models, *viz.* $p(x_k|x_{k-1}) = \mathcal{N}(x_k|x_{k-1}, \Delta)$, where $\mathcal{N}(.|\eta, \Delta)$ is a Gaussian distribution with mean η and covariance

$$\Delta = diag(\sigma_x^2, \sigma_y^2, \sigma_z^2, \sigma_{ax}^2, \sigma_{ay}^2, \sigma_{az}^2, \sigma_\theta^2, \sigma_\phi^2, \sigma_\psi^2).$$

In order to evaluate our 3D particles after their generation, we have to project them on the stereo images. The corresponding ellipses are obtained by a common quadric projection (Menezes *et al.*, 2005).

Let us now characterize both importance and measurement functions involved in our tracker.

3.1.1. Measurement function Our measurement function is based on skin color probability images and is inspired from the Thayananthan method (Thayananthan *et al.*, 2003).

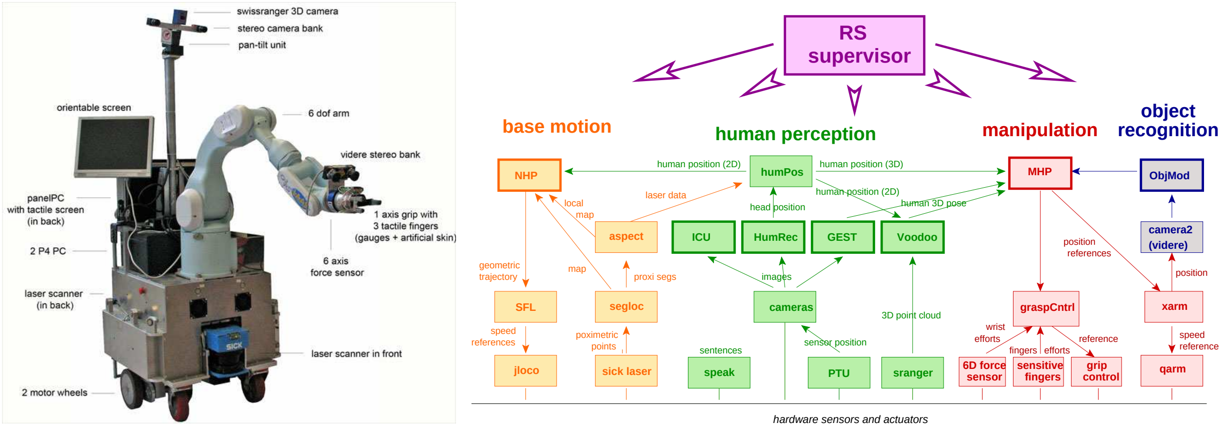
Each ellipse e - which is a projection of one particle - is given a likelihood $p(z, e)$ that depends on the average of skin color probabilities around the template corresponding to e. The pixels in the image are partitioned into a set of object pixels O, and a set of background pixels B. Assuming pixel-wise independence, the likelihood can be factored as

$$p(z, e) = \prod_{o \in O} (p(Ps(o), e)) \times \prod_{b \in B} (1 - p(Ps(b), e))$$

where $Ps(k)$ is the skin color probability at pixel location k.

The likelihood of the particle x is given by the merge of the two corresponding projected ellipses likelihood.

3.1.2. Importance function The importance function $\pi(.)$ is defined by a Gaussian mixture from the



(a)

(b)

Fig. 1. The Jido robot (a) and its software modular architecture (b).

triangulated 3D skin blobs. Let N be the number of detected 3D skin blobs and

$$b_i = (x_i, y_i, z_i, ax_i, ay_i, az_i, \theta_i, \phi_i, \psi_i), i \in \{1..N\}$$

the ellipsoid descriptions corresponding to each such region. An importance function $\pi(.)$ at location $\mathbf{x} = (x, y, z, ax, ay, az, \theta, \phi, \psi)$ follows, as the Gaussian mixture proposal

$$\pi(\mathbf{x}, z) = \sum_{i=1}^{N} \frac{1}{N} \mathcal{N}(\mathbf{x}|b_i, \Delta).$$

3.2 The ICU module

This module aims to track human upper-body (torso and head) in the video stream in order for the robot to be sure that it can interact with the person in front of it (recognition of its tutors) and to be able to follow somebody. The ICU module is composed of three main parts : (i) the face detector to found all faces in the video, (ii) the face recognition and (iii) the human upper-body tracker.

3.2.1. User face detector This detector is based on the well-known window scanning technique introduced in (Viola and Jones, 2001). This classifier covers a range of ±45° out-of-plane rotation of the user's face. It is used to switch between different modalities, but also to feed the face recognition part of the module.

3.2.2. User face recognition This part is based on the eigenface as described in (Turk and Pentland, 1991). It aims to classify facial regions $\mathcal{F}$ segmented from the face detector into either one class C_t out of the set $\{C_t\}_{1 \leq t \leq M}$ of M tutors of the robot. Each new detected face $\mathcal{F}(j)$, written as a $nm \times 1$ vector, is reconstructed in $\mathcal{F}_{r,t}$ by projecting it into $B(C_t)$. $\mathcal{F}$ is linked to the class C_t by its error norm:

$$\mathscr{D}(C_t|\mathcal{F}) = \frac{1}{n \times m} \sum_{j=1}^{n \times m} ((\mathcal{F}(j) - \mathcal{F}_{r,t}(j)) - \mu)^2,$$

where $\mathcal{F} - \mathcal{F}_{r,t}$ is the difference image, given that $|\mathcal{F} - \mathcal{F}_{r,t}|$ terms the DFFS [1], and μ the mean of $\mathcal{F} - \mathcal{F}_{r,t}$, and its associated likelihood

$$\mathscr{L}(C_t|\mathcal{F}) = \mathcal{N}(\mathscr{D}(C_t|\mathcal{F}); 0, \sigma_t)$$

where σ_t terms the standard deviation of distances of $B(C_t)$ training set.

The aforementioned likelihood $\mathscr{L}$ have to be thresholded in order to match the input face $\mathcal{F}$ with an already learned individual C_t. This threshold τ is deduced by computing likelihoods $\mathscr{L}$ between test image database with their own class C_t but also with the other classes noted $\neg C_t$.

Moreover, we investigated in preprocessing methods (Heseltine *et al.*, 2002) to improve the recognition accuracy. ROC curves have been generate from a database of 6000 faces to select the most meaningful image preprocessing. Histogram equalization is shown to outperform the other techniques for our database.

For a set of M learned tutors (classes) noted $\{C_t\}_{1 \leq t \leq M}$ and a detected face $\mathcal{F}$, we can define for each class C_t, the likelihood $\mathscr{L}_t = \mathscr{L}(\mathcal{F}, C_t)$ and an *a priori* probability $P(C_t|\mathcal{F})$ of labeling to C_t

$$\begin{cases} P(C_\emptyset|\mathcal{F}) = 1 \text{ and } \forall l \; P(C_t|\mathcal{F}) = 0 \text{ when } \forall l \; \mathscr{L}_t < \tau \\ P(C_\emptyset|\mathcal{F}) = 0 \text{ and } \forall l \; P(C_t|\mathcal{F}) = \frac{\mathscr{L}(C_t|\mathcal{F})}{\sum\limits_p \mathscr{L}(C_p|\mathcal{F})} \text{ otherwise} \end{cases}$$

where $C_\emptyset$ refers the void class.

3.2.3. User Tracking The tracking part is based on the I-Condensation algorithm also used in

[1] Distance From Face Space

Gest. The followed template is fit with its location $[u_k, v_k]'$, and its scale s_k, so that $\mathbf{x}_k = [u_k, v_k, s_k]$. In our human upper-body tracker, we consider multi-patches of distinct color distribution related to the head and the torso (Figure 2).

Moreover, taking into account the recognition step, the importance function related to the tracked class C_l becomes, with $\mathbf{b}_j$ the centroid of the j^{th} extracted face

$$\pi(\mathbf{x}_k^{(i)}|z_k) = \sum_{j=1}^{N_B} P(C_l|\mathcal{F}_j).\mathcal{N}(\mathbf{b}_j|\bar{X_B}, \Sigma_B),$$

where vector $\bar{X_B}$ and matrix Σ_B, which respectively term the mean and covariance of the offset from the ROI position to the centroid of the associated contour describing the face, are learned off-line.

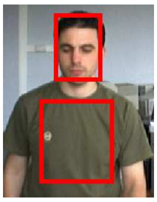


Fig. 2. The template.

3.3 The HumPos module

Evolved from work to develop experiments that are explained in (Sisbot *et al.*, 2006), the HumPos module, mainly based on laser data, detects elements that could be human legs. This module provides detection and tracking functionalities. The detection phase is divided in three stages: the first two stages are in charge of detecting legs by laser data and the last one is the correspondence with data obtained from ICU module (Figure 3).

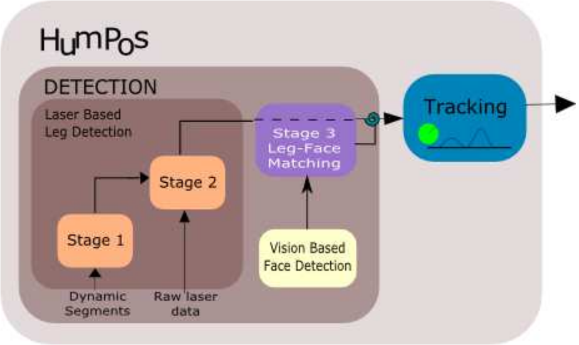


Fig. 3. The inner workings of Human Detection module, HumPos.

3.3.1. Detection At the first stage of the human detection, HumPos filters dynamic *segments* created from laser data (the ones which are not in the environment map) and, among them, keeps the ones which have a given size, form and proximity that could be considered as a leg. Once all legs are formed, pairs of legs are coupled. Each couple forms an item which will be included in the base list for human detection. It will be used in the second stage of the detection process.

In the second stage, HumPos creates blobs from a neighborhood of points obtained from *raw* laser data. Each of these blobs could represent legs depending on the quantity of points and on their distance from the laser sensor. Similar work is explained in (Xavier *et al.*, 2005) to detect cylinder and lines to find legs by analyzing their geometrical characteristics. As in the first step, legs are coupled and a list of items is obtained from this stage. This list is compared with the first list obtained in the previous stage and matched items increase their certainty.

Finally in this phase, at the third stage, a list of detected faces and position estimation obtained from the ICU module is compared with this list. This phase increases dramatically the confidence of matched items to be humans. In (Kleinehagenbrock *et al.*, 2002), a similar work is conducted by combining camera and laser to track people. The main difference resides in the camera system from the fact that it only detects faces to ensure the presence of human but does not estimate its real position in the space, obstructing in this manner multi-tracking.

3.3.2. Tracking In (Shulz *et al.*, 2001) or (Baba and Chatila, 2006), a particle filter dedicated to moving objects tracking with a laser scan is applied with good results, but we needed a lighter and simpler way to track items in order to satisfy real time constraints. In our case, we use a classical Kalman filter which takes in account multi dynamic hypotheses.

3.4 The Voodoo module

The Voodoo module is in charge of 3D human body tracking. It uses the swissranger device, which provides a 3D data point cloud of the scene. 178×144 depth images are acquired at 25 Hz.

The processing loop involves an *Iterative Closest Point* based algorithm which best fits a human model with the 3D data acquired from the swissranger camera. The model is constituted of 10 limbs, each of which is represented with a degenerated cylinder. The joints are modeled using some kind of “elastic bands” between limbs resulting in a 3 type classification: a universal joint with 3 DOF; a hinge joint with 1 DOF and 2 restricted DOF and finally an “elliptical joint” with 3 restricted joints. The curious reader can find a more complete description in (Knoop *et al.*, 2006).

The tracking can be initialized with the human position provided by the HumPos module. The initialization configuration is always supposed to be a standard one, that is to say human facing the camera with arms slightly moved aside from the body. The module can also integrate partially complete data about human pose from various sensor types, *e.g.* head or hand positions given by Gest or ICU modules.

4. RESULTS

4.1 Scenario

The main aim of the demonstration is to show a robot giving an object to a person. The object is supposed to be already grasped when the human arrives. This scenario is run under the control of a joint intention theory based supervisor described in (Clodic *et al.*, 2005).

The robot is navigating in an indoor environment. The person who wants to interact with the robot presents herself in front of it. The laser sensor is running and the HumPos module can thus detect a human presence at a high range distance. If someone is detected, the robot waits for the human to stop.

The Voodoo module is then initialized with data from HumPos and can track human body movements, in order to recognize some kind of interaction acceptation gesture from the human for example. Meanwhile, the ICU module tracks the global body position with the camera mounted on the pan-tilt unit. If face detections occur, they are compared to the tutor database. If the person standing in front of the robot belongs to learned tutors, the robot approaches and then the Gest module is launched. Human hands are tracked thanks to the stereo bank on the pan-tilt unit and their position is provided to modules in charge of the arm control to give an object to the human.

4.2 Experiments

The Gest module processes data at 5 Hz. This is enough to allow the tracked person to have natural movements. The precision of the hand position in depth lies between 10 and 20 cm, while the precision in x and y (width and height) is the cm. The lack of precision in depth is due to the proximity of the stereo cameras.

The integration of skin blobs segmentation on board of our JIDO robot, showed that its behavior is greatly influenced by the viewing condition changes in such a mobile robot context. Skin blobs segmentation must be used cautiously and should be coupled with other detection methods.

Snapshots of a typical sequence is shown on figure 4.

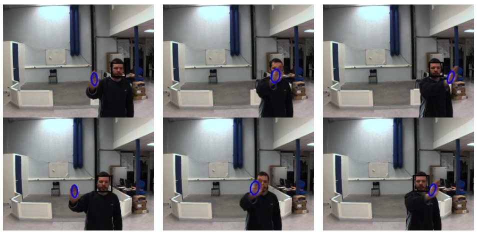


Fig. 4. Hand being tracked by the Gest module : the ellipses are the projection of the 3D state vector

The ICU module works between 10 and 15 Hz. It can easily detect face from 1m to 4.5m so, the tracker can work in this range of distance, even if the closer one can cause some difficulties due to the absence of the torso of the target. Figure 5 shows some snapshots of recognized tutors' faces where the detector marked – in red color – the detected faces but only those in green color are recognized from the previously learned faces, while Figure 6 involves occlusion of the target by another person crossing the field of view. The combination of multiple cues based likelihood and face recognition allows to keep track of the region of interest even after a complete occlusion.

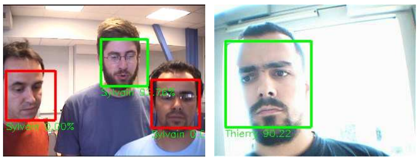


Fig. 5. Snapshots of detected/recognized faces with associated probabilities. The target is Sylvain (resp. Thierry) for the first (resp. last) frame.

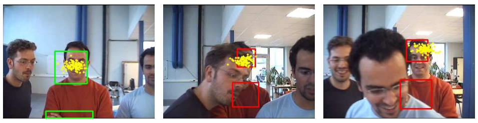


Fig. 6. Tracking scenario involving full occlusions between persons. Target recovery.

The Voodoo module processes data at 10 Hz on our robot software architecture, even if visualization is a bit slower. The tracker is efficient at a middle range distance. If the tracked person is too close, legs can possibly be out of the sensor field of view, what does not affect the accuracy of the upper body tracking. A screenshot of body configuration estimation is presented on figure 7.

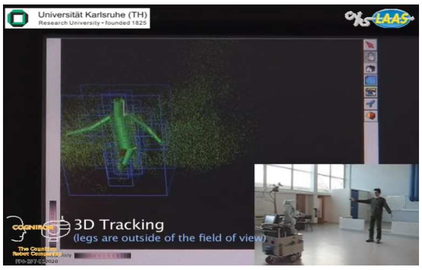

Universität Karlsruhe (TH)
Research University • founded 1825
LAAS
3D Tracking
(legs are outside of the field of view)

Fig. 7. The Voodoo module interface (background) showing the 3D body model estimated configuration with respect to the real scene (bottom right).

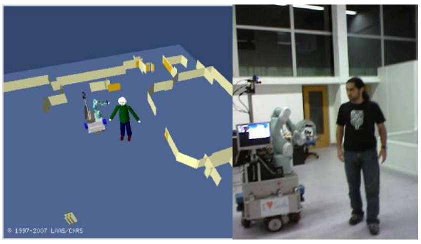


Fig. 8. The Jido robot environment representation (left) and the real human position (right).

The HumPos module processes data at 5 *Hz*. Thanks to laser data, the robot can localize itself and detect people in its environment (Figure 8).

5. CONCLUSION AND PERSPECTIVES

We have shown in this paper the result of an implementation of human perception algorithms on a robot. Different modalities are set up to localize and track the human at a wide range of distances. Face recognition is also performed to make the robot obey to offline learned tutors. These human perception abilities are used to give an object from the robot to the human.

Nevertheless, robustness can still be improved on some modules independently. We also plan to fuse more datas, for example *e.g.*, hand position tracked by the Gest module could be used to guide the Voodoo body tracker.Generally speaking, data fusion seems to be the key step in a robust human perception.

Some videos presenting parts of the scenario can be found at *www.laas.fr/∼mfontmar* and *www.laas.fr/∼tgerma*.

6. ACKNOWLEDGMENTS

We would especially like to thank all the people who worked on the robot, making it possible to achieve such results, among whom Aurélie Clodic, Sara Fleury, Matthieu Herrb, Akin Sisbot and Martin Lösch.

REFERENCES

Alami, R., R. Chatila, S. Fleury and F. Ingrand (1998). An architecture for autonomy. *Int. Journal of Robotic Research (IJRR'98)* **17**(4), 315–337.

Baba, Abedallatif and Raja Chatila (2006). Simultaneous environment mapping and mobile target tracking. *International Conference on Intelligent Autonomous Systems*.

Clodic, Aurélie, Vincent Montreuil, Rachid Alami and Raja Chatila (2005). A decisional framework for autonomous robots interacting with humans. In: *IEEE International Workshop on Robot and Human Interactive Communication (RO-MAN)*.

Heseltine, T, N. Pears and J. Austin (2002). Evaluation of image pre-processing techniques for eigenface based recognition. In: *Int. Conf. on Image and Graphics, SPIE*. pp. 677–685.

Isard, M. and A. Blake (1998). I-CONDENSATION: Unifying low-level and high-level tracking in a stochastic framework. In: *European Conf. on Computer Vision (ECCV'98)*. pp. 893–908.

Kleinehagenbrock, M., S. Lang, J. Fritsch, F. Lomker, G. A. Fink and G. Sagerer (2002). Person tracking with a mobile robot based on multi-modal anchoring. In: *Int. Workshop on Robot and Human Interactive Communication*. Berlin, Germany.

Knoop, Steffen, S. Vacek, K. Steinbach and R. Dillman (2006). Sensor fusion for model based 3d tracking. In: *Proceedings of International Conference on Multi-Sensor Fusion and Integration for Intelligent Systems (MFI)*. Heidelberg, Germany.

Menezes, P., F. Lerasle, J. Dias and R. Chatila (2005). A single camera motion capture system dedicated to gestures imitation. In: *Int. Conf. on Humanoid Robots (HUMANOID'05)*. Tsukuba. pp. 430–435.

Shulz, D., W. Burgard, D. Fox and A.B. Cremers (2001). Tracking multimple moving objects with a mobile robot. In: *Proc. of the IEEE Computer Society Conference on computer vision and pattern recognition (CVPR)*. Kauai,HW.

Sisbot, E. A., Clodic A., Marin L. F., Fontmarty M., Brethes L. and Alami R. (2006). Implementing a human-aware robot system. *15th IEEE International Symposium on robot and Human Interactive Comunication*.

Thayananthan, A., B. Stenger, P.H.S. Torr and R. Cipolla (2003). Learning a kinematic prior for tree-based filtering. In: *British Machine Vision Conf. (BMVC'03)*. Vol. 2. Norwick. pp. 589–598.

Turk, M.A. and A.P. Pentland (1991). Face recognition using eigenfaces. In: *Int. Conf. on Computer Vision and Pattern Recognition (CVPR'91)*. pp. 586–591.

Viola, P. and M. Jones (2001). Rapid Object Detection using a Boosted Cascade of Simple Features. In: *Int. Conf. on Computer Vision and Pattern Recognition (CVPR'01)*.

Xavier, Joao, Marco Pacheco, Daniel Castro and Antonio Ruano (2005). Last line, arc/circle and leg detection from laser scan data in a player driver. In: *In IEEE International Conference on Robotics and Automation*. Barcelona, Spain.