# Embodied Localization in Visually-guided Walk of Humanoid Robots

Hendry Ferreira Chame and Christine Chevallereau
*Robotics Team of the IRCCyN, Ecole Centrale de Nantes, CNRS, Nantes, France*

Keywords: Humanoid Robotics, Embodied Cognition, Ego-localization, Machine Vision.

Abstract: Humanoid robots are conceived to resemble the body and comportment of the human beings. Among the behavior repertoire, the possibility of executing visually-guided tasks is crucial for individual adaptation and relies on the on-board sensory system. However, the research on walk and localization is far from conclusive. Given the difficulties in the processing of the visual feedback, some studies have treated the problem by placing external sensors on the environment; thus neglecting the corporal metaphor. Others, despite exploring on-board solutions; have relied on an extensive model of the environment, thus considering the system as an information processing unit, abstracted from a body. This work presents a methodology to achieve embodied localization to serve visually-guided walk. The solution leans on robust segmentation from monocular vision, ego-cylindrical localization, and minimal knowledge about stimuli in the environment.

## 1 INTRODUCTION

Humanoid robots are conceived to resemble the body and the comportment of the human beings. Among the behavior repertoire, the possibility of executing visually-guided tasks is crucial for individual adaptation and relies on the on-board sensory system. However, the research on walk and localization is far from conclusive. When using vision, estimates on localization strongly depend on the quality of the object tracking process. Given the differences between the human and the artificial vision system, the recognition (or segmentation) of stimuli on the scene can be difficult; and different approaches are available. Furthermore, an interesting debate has been taking place on the way to represent the relation between the agent's posture, the objects' posture, and the desired locations in space. Traditional approaches has considered solutions where the agent is conceived as an information processing unit, in a decoupled Cartesianist mind-body view; where intelligent behavior is regarded as a symbolic manipulation process from actual sensory input (Shapiro, 2007). Recently, subjectivity in being considered in a more Heideggerian sense, where agency and interactive coping occupy center stage (Anderson, 2003). Such as, real-world thinking occurs in particular situations, and is employed for specific practical ends. Thereby, cognition is viewed as embodied or within a situated activity.

Starting from the notion of embodiment, the objective of this work is to explore the possibilities of designing solutions to the problem of visually-guided positioning in relation to stimuli on the environment. Thus, a methodology that leans on robust segmentation from monocular vision will be reported. The solution proposes ego-cylindrical localization, requiring of minimal knowledge about stimuli. The sections of this paper are organized as follows. Section 2 explores related works in the field of on-board localization. Given the difficulties of object tracking, Section 3 introduces the problem of image processing by exploring some techniques to achieve robust segmentation. The Markov Random Field (MRF) formalism will be discussed in more detail, due to the good results it provides on image segmentation. Section 4 formalizes the localization task via a case study. Section 5 presents the results obtained. Finally, conclusions and research perspectives are presented.

## 2 RELATED WORK

On-board visual localization relies on robust object tracking, which is a challenging task for walking robots. Indeed, certain difficulties have been reported when attempting to make use of the captured images, since walk introduces motion blur noise. Some studies have treated the problem by placing the visual system on the environment (e.g., (Lewis M.A. and Simo L.S., 1999)). Though, occlusions may compromise

Ferreira Chame H. and Chevallereau C..
Embodied Localization in Visually-guided Walk of Humanoid Robots.
DOI: 10.5220/0005063001650174
In *Proceedings of the 11th International Conference on Informatics in Control, Automation and Robotics* (ICINCO-2014), pages 165-174
ISBN: 978-989-758-040-6

the localization given the robot's motions. Alternatives include estimates in position through other sensor modalities like external microphones, which capture the robot's intrinsic noise (Allen et al., 2012). Unfortunately, the location of stimuli and the robot in relation to the sensors may also compromise the results. Furthermore, the orientation component cannot be estimated in this way. In general, exteroceptive solutions neglect the corporal metaphor, impose the condition that the environment must be adapted to the problem, and are very sensitive to modeling imprecisions.

When considering on-board solutions, the limited control over the head's direction and its effect over the visual output, has complicated the task (Michel et al., 2005). To treat this problem, the compensation for the head's motion by taking advantage of the kinematic and geometric models of the robot has been attempted; such as, a virtual camera has been defined to cancel the sway motion in the visual features for continuous visual servoing (Dune et al., 2010). Though, considerable delays may be involved in the vision processing; due to digital image treatment and video data transference from the on-board camera to the computer system (Moughlbay et al., 2013). These delays can restrict the applications of real-time visual servoing techniques in closed-loop. Furthermore, physiological evidence has also reported considerable delays in the human visuo-motor loop (Miall et al., 1993). The feedback is estimated to take around 130 ms for ocular-motor control and 110-150 ms for proprioceptive control. According to these figures, the performance observed in natural beings may be better explained by the organization and the efficiency in the management of the available resources, rather than by the computational power. In addition, continuous visual control during walking may not be necessary since depending on the walking stage, images have greater or less relevance for the localization (the head's motion may produce blurred images at certain points). So considerable processing overhead may be added with little benefit for localization.

The task representation has also been a topic of interest. The displacement to be accomplished has been referenced within a global map; that the agent may possess, update, or build while navigating. For example, in a work developed by (Hornung et al., 2010), the problem of indoor localization is tackled by adapting a range sensor to the robot's head. The posture of the robot is estimated within a known volumetric map of the environment; such as, the on-board measurements parametrize a probabilistic search routine. In a work developed by (Robert Cupec, 2005), a global localization policy is combined with local references to enhance the accuracy when stepping over small obstacles. The strategy is based on an interesting method for directing the gaze by maximizing the visual information; but evidences the limitations of the global localization approach where the accuracy is greatly affected and strongly depends on the quality of modeling (including parameters estimation), and the noise in the measurements. For both of these works, the localization-and-locomotion task has been modeled as a control problem, with the body playing the role of a mere tool that has to be commanded appropriately (Hoffmann and Pfeifer, 2012).

The discussion has exposed at this point some important aspects about localization in humanoid robotics. It has been assessed the reliability of the on-board sensory to effectively accomplish the task, given the noise introduced by motion. Furthermore, the role of the agent in relation to the environment has been investigated; in particular, the extent to which the environment must be known or adapted to the agent for the attainment of the localization task. Lastly, the convenience of using a global reference policy has been contrasted to locally referencing stimuli, in relation to the precision obtained for localization. This research starts from the hypothesis that on-board localization can be achieved by relying on robust object segmentation, with minimal knowledge about the environment, and defining a sensory egocentric reference system. In the following, these aspects will be discussed.

## 3 IMAGE SEGMENTATION

Image segmentation and object tracking are hard processes to achieve. In the literature, a huge number of techniques are available, where each one imposes certain constraints. An in-deep treatment of the topic cannot be accomplished here; thus, some of the explored proposals that showed good results are going to be briefly discussed.

The first approach considered was the classical k-means algorithm (MacQueen, 1967), which is a convenient technique to obtain clusters. The method is not very efficient for real-time applications given its high computational complexity $\varsigma = O(n^{dk+1}\log(n))$, for $d$ dimension feature vectors, $k$ clusters and $n$ elements (pixels in the case of images). Also, $k$ is required which significantly constraints the characteristics of the images to be treated. The expectation maximization (EM) algorithm (Dempster et al., 1977) is more efficient and general, in the sense that the clusters are represented by probability distributions and not just by the means. Unfortunately, it also requires of $k$ as an input parameter.

Another technique explored was the continuously adaptive mean shift (CAMShift) algorithm, which performs color-based tracking. CAMShift essentially climbs the gradient of a back-projected probability distribution from a color histogram, and finds the nearest peak within a search window. When the camera is fixed, the algorithm offers reasonably good results. However, for on-board tracking, the motions of the camera affect the color distribution of the object since variations in the point of view result in changes in illumination; which degrades the obtained segmentation. Improvements have been proposed by (Exner et al., 2010) and consist in the accumulation of multiple histograms.

As alternative to color-based tracking, feature-based techniques have also been explored. The differential Lucas-Kanade method (Lucas and Kanade, 1981) estimates the optical flow by using a least squares criterion. The algorithm assumes brightness constancy, spatial coherence, and small displacement of the features between frames (high frame rate). The last requirement makes the technique unsuited to systems operating at low frequencies, so the tracking may be lost due to the constrained local search.

Some techniques have focused, by other hand, on the pre-processing of the image to reduce the motion blurs introduced by the walk (Pretto et al., 2009). The observed image $b(x,y) = h(x,y) * f(x,y) + n(x,y)$ is the result of the convolution operation of the a blurring function $h(x,y)$, also known as the point spread function (PSF); over the original image $f(x,y)$ and the added noise $n(x,y)$. The goal is to restore $f(x,y)$ from an estimation of $h(x.y)$. Several techniques have been proposed to deconvolve $f(x,y)$ (e.g., the Richardson-Lucy algorithm and Wiener filter). Unfortunately, it is not so simple to estimate the PSF for random motions, and the quality of the results strongly depend on it.

In general, for most of the techniques discussed, the elements under analysis have been individual pixels of the image; which becomes a main drawback in the presence of noise. Spacial tracking techniques generally aim to match points of interest in a sequence of images; by assuming that the time interval between frames is small enough to perform a local search at low computational cost (generally around a neighborhood). Unfortunately, to ensure real-time responsiveness; the calculation of the features tend to be simple in order to be fast. This conditions the robustness under disturbances like the motion blurs. Besides, similarly to the delays observed in natural beings; certain platforms (e.g. the Nao robot) cannot afford a high frame-rate to satisfy the tracking conditions.

## 3.1 Markof Random Fields (MRF)

An alternative to spacial tracking is to assume no relation between successive images; such as, the segmentation can be achieved by only relying on the color model of the object. The MRF formalism considers the spacial coherence between regions of pixels on the image; it is an interesting approach to obtain robust segmentation and is going to be discussed here.

The observed image $F = \{f_s \mid s \in I\}$ consists of the spectral component values registered in a color-space $\eta$ at which each pixel $s$ is denoted by the vector $f_s$. The label of interest $\hat{\varphi}$ is the one that maximizes the a posteriori probability $P(\varphi \mid F)$:

$$argmax_{\varphi\in\Phi}\prod_{s\in I} P(f_s \mid \varphi_s)P(\varphi), \qquad (1)$$

where $\Phi$ denotes the set of all possible labellings. Since the goal is to segment the image into homogeneous regions, a pixel class $\lambda$ should correspond to one or more homogeneous color patches in the input image. Such regularities can be modeled by an additive white noise with covariance $\Sigma_\lambda$ centered around the expected color value $\mu_\lambda$. Thus, $P(f_s \mid \varphi_s)$ follows a Gaussian distribution and pixel classes $\lambda \in \Lambda = \{1, 2, ...L\}$ are represented by the mean vectors $\mu_\lambda$ and the covariance matrices $\Sigma_\lambda$. Furthermore, $P(\varphi)$ is a MRF with respect to a first order neighborhood system (as shown in Fig. 1).

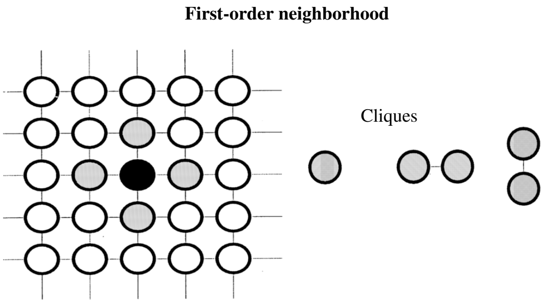


Figure 1: First-order neighborhood system. Single pixel cliques are called singletons, horizontal and vertical cliques are called doubletons (Kato et al., 2001).

According to the *Hammersley-Clifford theorem*, $P(*)$ follows a Gibbs distribution:

$$P(\varphi) = \frac{e^{-U(\varphi)}}{Z(\gamma)} = \frac{\prod_{c\in C} e^{-V_c(\varphi_c)}}{Z(\gamma)}, \qquad (2)$$

where $U(\varphi)$ is called an *energy function*, $Z(\gamma) = \sum_{\varphi\in\Phi} e^{-U(\varphi)}$ is the normalizing constant (or partition function) and $V_c$ denotes the *clique potential* of clique $c \in C$ having the label configuration $\varphi_c$. $C$ is the set of spatial second order cliques (i.e., doubletons). The energies of singletons directly reflect the probabilistic modeling of labels without context, while doubleton clique potentials express the relationship between neighboring pixel labels. The energy function denoted on the MRF image segmentation model, as proposed by (Kato et al., 2001); has the following form:

$$U(\varphi,F) = \sum_{s\in I}(\ln(g)+\frac{k}{2}) + \beta \sum_{\{s,r\}\in C} \delta(\varphi_s,\varphi_r), \quad (3)$$

where $g = \sqrt{(2\pi)^3|\Sigma_{\varphi_s}|}$, $k = (f_s - \mu_{\varphi_s})\Sigma_{\varphi_s}{}^{-1}(f_s - \mu_{\varphi_s})^t$; and $\delta(\varphi_s,\varphi_r) = 1$ if $\varphi_s \neq \varphi_r$ and zero otherwise. The parameter $\beta > 0$ controls the homogeneity of the regions; as it increases, the regions become more homogeneous. The function $U(\varphi,F)$ is non-convex, so the convergence to the global optimum cannot be ensured since the calculation of $Z(\gamma)$ in (2) is intractable. In practice, combinatorial optimization techniques such as iterated conditional modes (ICM) (Besag, 1986) are employed to achieve the segmentation. The next state $\hat{\varphi}_s^{k+1}$ is determined by

$$\hat{\varphi}_s^{k+1} \leftarrow \arg \min_{\varphi_{s\in\{1,\ldots,L\}}} U(\hat{\varphi}^k,F). \quad (4)$$

The stop condition is attained when

$$\hat{\varphi}_s^{k+1} = \hat{\varphi}_s^k, \forall s. \quad (5)$$

To summarize, the parameters of the system are $\Theta = (\mu_\lambda, \Sigma_\lambda, \beta)$. In case when $\Theta$ is provided by the user, a *supervised* segmentation is obtained. Otherwise, $\Theta$ must be automatically estimated simultaneously to $\varphi$, which is known as *unsupervised* segmentation. For the later case, the role of the parameter $\beta$ may vary in time, as pointed out by (Deng and Clausi, 2004).

# 4 ON-BOARD LOCALIZATION

The object is assumed to be previously known with its dimension. The robot has to localize it, thus the relation in position and orientation with respect to the robot has to be defined. The parameters $\Theta = (\xi, \psi)$ of the system are the geometrical properties of the object $\xi$, and its color model $\psi$.

## 4.1 Sensory Ego-cylinder

The robot is assumed to be walking on a plane, such as the movable frame $B$ is fixed to the ground. The origin of the frame corresponds to the intermediate point between the center of projection of both feet on the ground (see Fig. 2). Analogously, $B_x$ is the mean direction between the major orientation axis of each foot. The axis $B_z$ is chosen to be perpendicular to the ground plane, and $B_y \perp B_x$.

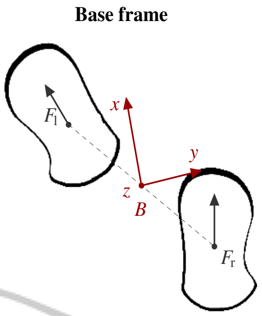


Figure 2: Representation of the base frame $B$.

The concept of ego-sphere, as presented in (Bodiroza et al., 2011), is an interesting proposal to express the ego-localization of the saliency of stimuli on the environment. Although, the cylindrical geometry seems to be more appealing to our case, given that it is simpler and convenient to represent the positions of objects moving on a plane. Thus, we employ an ego-cylinder principle for localization (as shown in Fig. 3). In relation to the orientation component, only the azimuth $\phi$ around $B_z$ can be corrected by the walk primitives of the robot; so the ego-cylinder is extended to include the magnitude of $\phi$ as follows

$$P = \begin{bmatrix} \rho & \theta & z & \phi \end{bmatrix}^t. \quad (6)$$

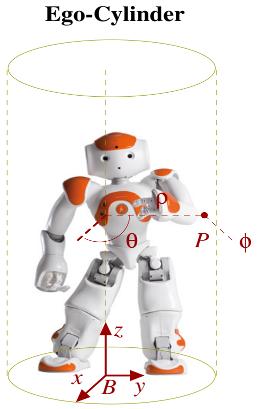


Figure 3: Representation of the ego-cylinder localization. In the image, $B$ corresponds to the base frame, and $P$ represents the localization of an object in the environment. The magnitude of the rotation $\phi$ around $B_z$ is represented by the direction emerging from the cylinder's surface.

## 4.2 Localization

The localization of the object in the scene is based on the definition of four frames, as depicted in Fig. 4. The pose $P$ of the object can be known with respect to the base frame $B$ through the definition of the homogeneous transformation matrix

$${}^{\mathrm{B}}T_{\mathrm{O}} = {}^{\mathrm{B}}T_{\mathrm{H}}(q)\,{}^{\mathrm{H}}T_{\mathrm{C}}\,{}^{\mathrm{C}}T_{\mathrm{O}}, \tag{7}$$

where the transformation ${}^{\mathrm{B}}T_{\mathrm{H}}(q)$ expresses the head frame $H$ in the base frame $B$, and depends on the actual joint configuration $q$ of the robot. The transformation ${}^{\mathrm{H}}T_{\mathrm{C}}$ is constant and expresses the camera frame $C$ in frame $H$. The transformation ${}^{\mathrm{C}}T_{\mathrm{O}}$ expresses the object frame $O$ in frame $C$, and is determined from the 3D pose

$${}^{\mathrm{C}}O = [\zeta \quad \omega]^{\mathrm{t}} = [[x \quad y \quad z] \quad [\gamma \quad \beta \quad \theta]]^{\mathrm{t}}, \tag{8}$$

where $\zeta$ is the position component and $\omega$ is the orientation component. The calculation of ${}^{\mathrm{C}}O$ depends on the geometry of the object model, some examples are given as case studies in Section 4.4.

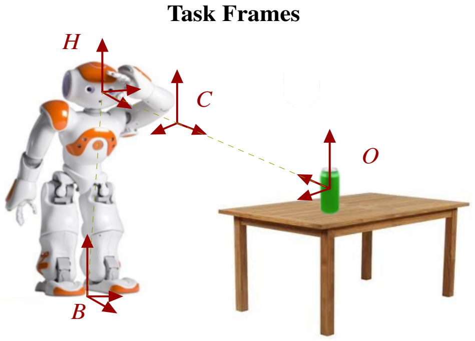


Figure 4: Definition of the reference frames to solve the localization task. In the image, $B$ corresponds to the base frame, $H$ to the head frame, $C$ to the camera frame, and $O$ to the object frame.

The transformation ${}^{B}T_{\mathrm{O}}$ allows the definition of the localization of the object in the ego-cylinder by expressing the position of the center of frame $O$ in cylindrical coordinates, and adding the orientation $\phi$ of the object around $B_z$ as defined in (6).

## 4.3 Visually-guided Walk

Since the base frame $B$ is mobile, the transformation ${}^{\mathrm{B}}T_{\mathrm{B}^*}$ between its current location and the desired location $B^*$ in relation to the object is given by

$${}^{\mathrm{B}}T_{\mathrm{B}^*} = {}^{\mathrm{B}}T_{\mathrm{O}}\,{}^{\mathrm{O}}T_{\mathrm{B}^*}, \tag{9}$$

where the transformation ${}^{\mathrm{O}}T_{\mathrm{B}^*}$ is defined by demonstration. In other words, by placing the robot at the desired pose in relation to the object.

A difference in location ${}^{\mathrm{B}}d$ between the current and the desired configuration, presents the same structure of (6), and is obtained from ${}^{\mathrm{B}}T_{\mathrm{B}^*}$. The first three components are determined by expressing the position component of the transformation in cylindrical coordinates; whereas the four coordinate is extracted from the rotational component of ${}^{\mathrm{B}}T_{\mathrm{B}^*}$ and corresponds to the rotation around $B_z$. A direction of motion $\bar{M}$ can be determined from ${}^{\mathrm{B}}d$ as follows

$${}^{\mathrm{B}}\bar{M} = sat({}^{\mathrm{B}}d,\lambda), \tag{10}$$

where $sat$ is a saturation function for the position and orientation components of ${}^{\mathrm{B}}d$, and $\lambda$ are the corresponding thresholds.

## 4.4 Case Studies

Object localization based on visual tracking, is greatly dependent on the quality of the segmentation process. If the later is successful, simpler geometrical models are enough to accomplish the task. Thus, instead of fitting rich 3D meshes to images (such as in (Legrand et al., 2002)); simple geometrical containers were considered as models. The idea behind this philosophy is defining reusable models acting as oriented wrappers to objects of potential interest on the scene. Next, the modeling of two of these containers and how to estimate them from the image blob is going to be detailed.

### 4.4.1 Cylindrical Wrapper

The frame $O$ is attached to the center of mass of the model as shown in Fig. 5. As a result, ${}^{\mathrm{C}}O$ in (8) is estimated. Given the symmetry of the shape, the projection of the object in the image plane is not affected by the rotation $\beta$ around $O_y$; so it is assumed to be constant.

**Depth Estimation.** The blob is approximately centered on the image to avoid calculations over a clipped projection of the object. In order to estimate the position of frame $O$ (as illustrated in Fig. 6); the depth of $L$ and $R$ with respect to the frame $C$ must be calculated through the function

$$d(r,r',f) = \sqrt{\left(\frac{r}{\sin(\gamma)}\right)^2 + r^2}, \tag{11}$$

where $r$ is the radius of the cylinder, $r'$ is its projection on the image plane, $\gamma = \mathrm{atan2}(r',f)$, and $f$ is the focal length of the camera. ${}^{\mathrm{C}}L_z$ and ${}^{\mathrm{C}}R_z$ are calculated

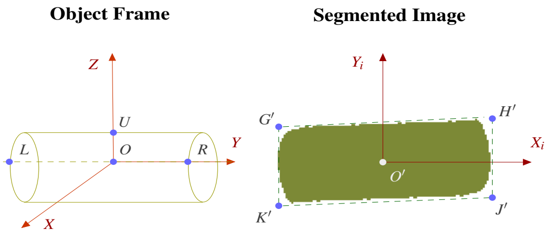
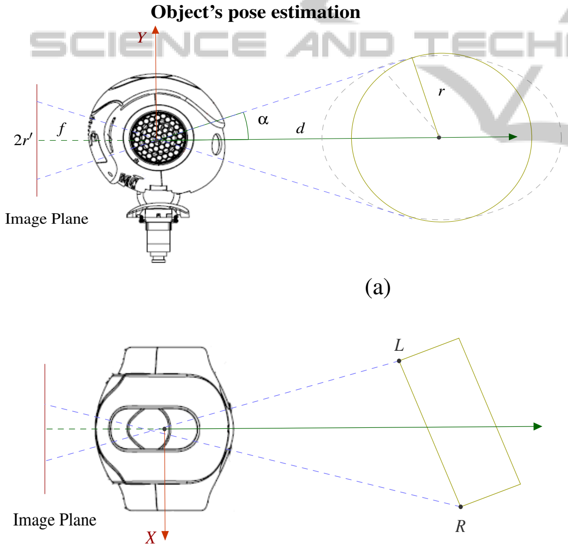


Figure 5: Definition of the cylindrical object model. On the left, the 3D representation of the object frame $O$ and the definition of four points of interest. On the right, the illustration of the segmented blob and the definition of image features from the oriented bounding box.

by tacking $r'$ to be $|G' - K'|/2$ and $|H' - J'|/2$ respectively. The analytical expression will be exact if the orientation component ${}^{C}O_{\theta} = 0$ in (8); otherwise inaccuracy will be introduced.



(a)

(b)

Figure 6: Estimation of the object's depth. a) The model assumes ${}^{C}P_{\phi} = 0$, b) XZ visualization of the scenario where the circumference corresponds to an ellipse and the distance from the projective ray and the center $O$ is larger than $r$.

**Position Estimation.** The position of a 3D point $X$ in frame $C$ can be calculated by the definition of the projective function ${}^{C}X = p(X_z, X', C', f)$, such as

$$p(X_z, X', C', f) = \begin{bmatrix} ((X'_x - C'_x)X_z)/f \\ ((X'_y - C'_y)X_z)/f \\ X_z \end{bmatrix}. \tag{12}$$

Where $X'$ is the projection of $X$ in the image plane, and $C'$ is the image center. Thus, the position component ${}^{C}O_{\zeta}$ in (8) is given by

$${}^{C}O_{\zeta} = \begin{bmatrix} M_x & M_y & M_z + r \end{bmatrix}^{t}, \tag{13}$$

where $M = \text{mean}({}^{C}L, {}^{C}R)$. The other features of the object model are calculated using (12) such as

$$\begin{bmatrix} {}^{C}L \\ {}^{C}R \\ {}^{C}U \end{bmatrix} = \begin{bmatrix} p({}^{C}L_z, \text{mean}(G', K'), C', f) \\ p({}^{C}R_z, \text{mean}(H', J'), C', f) \\ p({}^{C}O_z, \text{mean}(G', H'), C', f) \end{bmatrix}. \tag{14}$$

**Orientation Estimation.** The orientation component ${}^{C}O_{\omega}$ in (8) is obtained from the relation between ${}^{C}R$, ${}^{C}L$, ${}^{C}U$, and ${}^{C}O$. It is extracted from the rotation matrix

$${}^{C}R = \begin{bmatrix} s & n & a \end{bmatrix} = \begin{bmatrix} \hat{H} & \hat{V} & (\hat{H} \times \hat{V}) \end{bmatrix}, \tag{15}$$

with $\hat{H} = ({}^{C}R - {}^{C}L)/|{}^{C}R - {}^{C}L|$, and $\hat{V} = ({}^{C}U - {}^{C}O)/|{}^{C}U - {}^{C}O|$.

### 4.4.2 Rectangular Surface

Rectangles are useful geometric models for tracking surfaces in walls, doors and furnitures (e.g., drawers). The model is simpler than the previous one, since the vertical axis $O_z$ is assumed to be perpendicular to the ground. The points defining $O$ correspond to those of Fig. 5, except that $R$ and $L$ are now contained in the ZY plane. The features tracked in the image are also similar to the prior case. However, the calculation for the depth of $O$ changes to be given by

$$d(h, h', f) = \frac{hf}{h'}, \tag{16}$$

where $f$ is the focal distance of the camera, $h$ is half of the height of the rectangle, and $h'$ is the image projection of $h$. The relation between the image features and the location of ${}^{C}R$, ${}^{C}L$, and ${}^{C}U$ is similar to the previous case; though, ${}^{C}O = \text{mean}({}^{C}L, {}^{C}R)$.

## 5 RESULTS

The study has been conducted in three stages. At first, several segmentation algorithms were explored to assess the robustness against the motions of the camera. Next, a localization task has been simulated to optimize the development of the algorithms and to detect errors. Lastly, after obtaining a stable and correct execution, the program has been evaluated in the experimental platform. In the following, the results obtained at each of these stages will be reported.

## 5.1 The Object Tracking Algorithm

The object tracking program was implemented in the C++ language and included the OpenCV 2.4.8 library. The routine presented the structure shown in Algorithm 1. In relation to (3), the *initializeColorModel* method estimates the statistical parameters ($\mu$ and $\Sigma$) of the color model from a region labeled by the user. No color model is required for the background, since the appearance of new objects in the scene would affect the tracking. Given the variations in lighting, $n$ sampled frames are averaged to reduce the noise (usually $n = 10$ gives good results); and then provided to the initialization routine. The *objectIsCentered* method rotates the head until the segmented blob is approximately centered on the image. Once accomplished, *doFeatureExtration* calculates the object's pose in the camera frame $C$.

**Algorithm 1:** Object Tracking.

```
procedure DOTRACKING
    $initializeColorModel()$
    while $run$ do
        while $objectIsCentered() = false$ do
            $doSegmentation()$
        $doFeatureExtration()$
```

The segmentation routine consisted in a customization of the MRF *supervised* technique (see Algorithm 2). The algorithm possesses a computational complexity $\varsigma = O(n^{|\Phi|})$, where $n$ is the number of $s$ pixels of the image $I$, as described in Section 3.1. Since the problem was to recognize the object from the background, $|\Phi| = 2$. In addition, the images were processed in the YUV color-space, thus, $|\eta| = 3$. The *localEnergy* function corresponds to (3) with the difference that it is calculated with the color model of the object (without the background model as explained before). The *initialize* method proposes an initial segmentation candidate $\hat{\varphi}$ by minimizing the singleton term $\sum_{s \in S} (\ln(g) + \frac{k}{2})$. The evaluation of the segmentation algorithm has shown that, for naturally illuminated scenes (see Fig. 7), it is quite robust under camera motions when detecting colored objects with diffuse, non-specular reflective textures (as illustrated in Fig. 8). The objects don't have to possess uniform or single colors as depicted in Fig. 9. For the case of artificially illuminated scenes, in particular, under low-frequency lighting; more samplings may be required to estimate the color model.

**Algorithm 2:** Segmentation.

```
procedure DOSEGMENTATION
    $\hat{\varphi}(i,j) \leftarrow Initialize()$    ▷ Singleton initialization
    $e_{Old} \leftarrow 0$
    repeat
        $e \leftarrow 0$
        for $i = 0 \rightarrow i < height$ do
            $min_e \leftarrow localEnergy(i,j,\hat{\varphi}(i,j))$
            for $j = 0 \rightarrow j < width$ do
                for $\lambda = 0 \rightarrow \lambda < |\Phi|$ do
                    $c_e \leftarrow localEnergy(i,j,\lambda)$    ▷ current energy
                    if $c_e < min_e$ then
                        $\hat{\varphi}(i,j) \leftarrow \lambda$
                        $min_e \leftarrow c_e$
            $e \leftarrow e + min_e$
        $\Delta e \leftarrow abs(e_{Old} - e)$
        $e_{Old} \leftarrow e$    ▷ stop when the change is too small
    until $\Delta e > t$
```

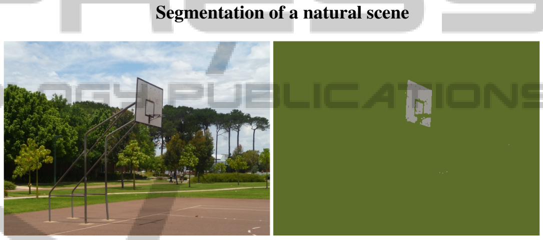


Figure 7: Segmentation of a natural scene. On the left, the original image where a color sample was taken from the white backboard. On the right, the segmentation achieved.

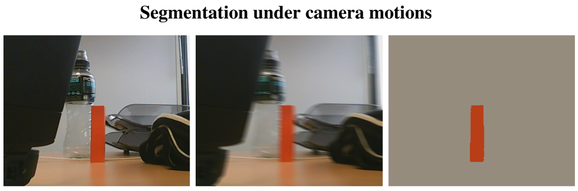


Figure 8: Segmentation under camera motions. On the left, the still image of the scene. In the center, a random motion was applied to the camera. On the right, the segmentation obtained.

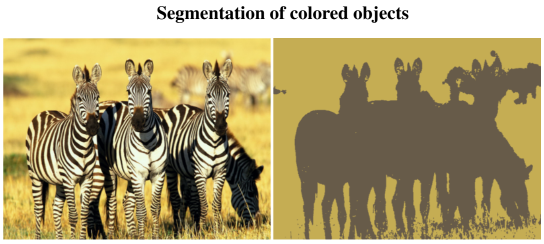


Figure 9: Segmentation of colored objects. On the left, the original image of a group of zebras. On the right, the segmentation achieved.

## 5.2 The Simulation Environment

The designed methodology can serve at two distinct objectives. The first one is a typical information-processing scenario, that employs the egocentric localization to guide the robot to relative coordinates in the scene (e.g., requiring it to be at 20 cm in front of the object). The precision of this task will be affected by the errors introduced in the image projection, and the approximations of the object model. The second one is to show the robot, by demonstration, how it has to be placed with respect to the target, such as, the perception would be embodied. Here, it is not so important the absolute precision of the estimates, but the the way the robot perceives its body in relation to the stimulous.

In order to assess the performance under the imprecisions described; a simulated environment has been designed in Webots 7.0.4. In the conceived scenario, the object of interest corresponded to a red soda can placed over a table (as illustrated in Fig. 10). The desired configuration ${}^{B^*}T_O$ was specified by positioning the robot in front of the can with its body oriented at $\phi \approx \pi$ with respect to it.

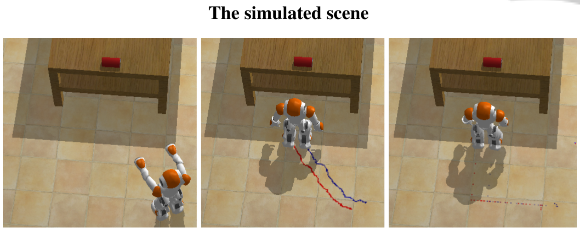


Figure 10: The approach task modeled in Webots. On the left, the robot's original pose. In the center, the followed trajectory. On the right, the desired pose with respect to the red can on the table. As it can be seen, despite the modeling errors, the robot was able to converge to a location very similar to the demonstrated one.

For testing the localization task, since the walk routines cannot act on the $z$ component of (10); the motion vector ${}^{B}M' = [M_\rho, M_\theta, M_\phi]^t$ was given as the direction of displacement to the robot's walk primitive. No trajectory generation nor control was considered. Figure 11 illustrates the evolution of the localization along the followed trajectory. Figure 12 compares the on-board estimations with the measurements provided by Webots. Despite the initial estimations are not very precise, as the robot approached the target, the precision increased enough to allow it to convergence to the desired pose.

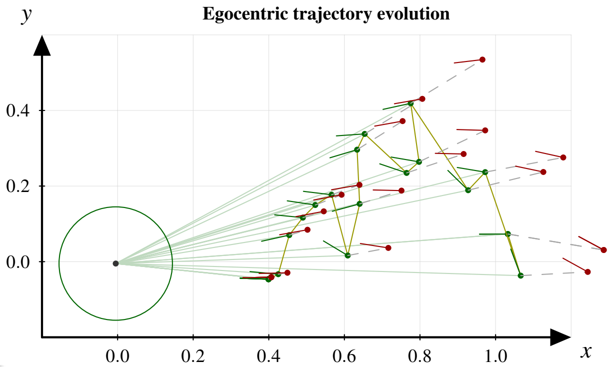


Figure 11: XY egocentric visualization of the localization as perceived in $B$. The circumference represents the egocylinder. In red the real values, in green the estimations. Distances are expressed in m.

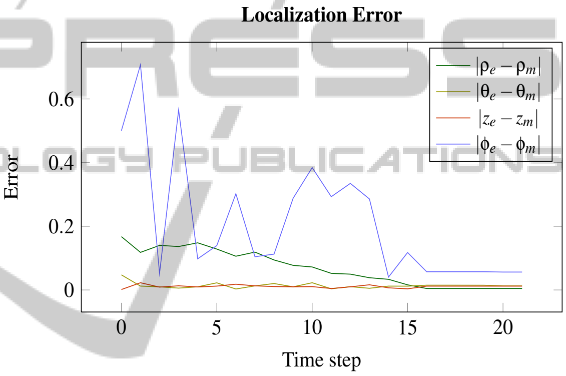


Figure 12: Evolution of the localization error between the estimations $e$ and the measurements $m$.

## 5.3 The Experimental Environment

In the experimental task the robot was placed in an unstructured scene (the robotic lab). It was required to approach a planar yellow rectangle (whose model was detailed in Section 4.4.2); to a relative pose captured by demonstration. As depicted in Fig.13, there are different sorts of colored stimuli on the environment. Despite this variability, the robot was able to converge to the desired pose.

# 6 CONCLUSIONS AND FUTURE WORK

This study has explored the possibility of obtaining embodied visual localization to serve humanoid robotics walk. For this purpose, a method based on monocular vision was developed. Given the noise in the on-board measurements, the research proposed to verify that a sequential look-then-move policy would

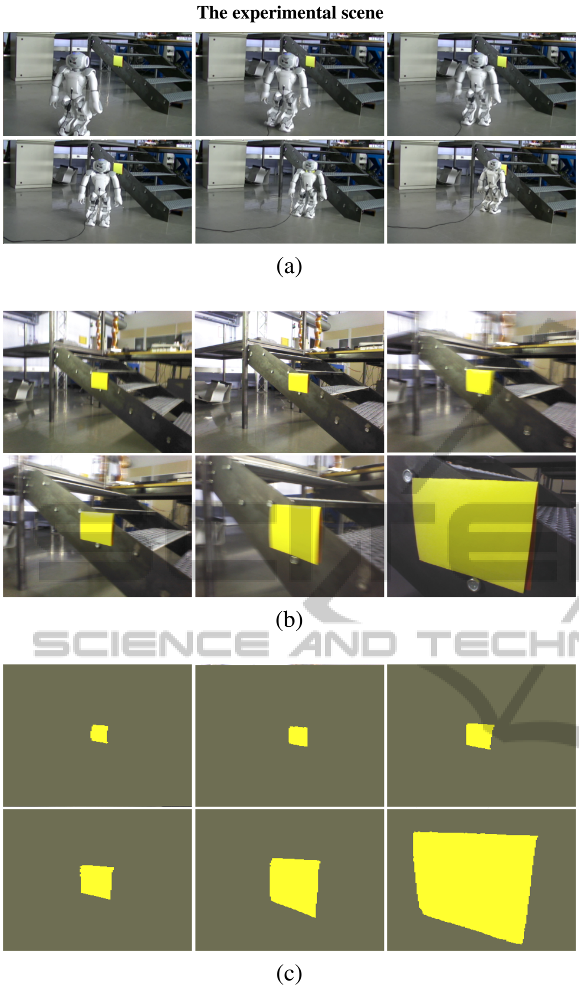


Figure 13: The experimental environment. a) The scene captured from an external camera, b) The on-board view, c) The segmentations obtained.

be sufficient to perform the task; such as, computational efforts could be invested in the achievement of robust object recognition. In this sense, the MRF formalism proved to be a convenient framework to define the problem of color-based, supervised, image segmentation under motion noise. The ego-localization representation involved the definition of a perceptive ego-cylinder. Case studies has been proposed to illustrate the philosophy behind the methodology, and consisted in the definition of simple and reusable models to wrap objects in the environment. Simulations and experimentations were conducted and have shown that, despite the simplicity of the models and the perturbations involved, the robot was able to converge to a desired pose in relation to the object by relying exclusively on local estimates. The results obtained addressed the benefits of embodiment for perception and cognition in robotics, as compared to the information-processing paradigm.

On-going efforts are aiming at including a top-down feature attention mechanism for assisting tracking when objects leave the field of vision. Futhermore, the error in position and orientation was independently regulated, thus resulting in holonomic motions. The obtained trajectories can be improved by defining a non-holonomic, ego-centric, trajectory generation policy; which is currently under study and will endorse the agent with a more human walk style.

## ACKNOWLEDGEMENTS

This research has been funded by the Ecole Centrale de Nantes; and the CAPES Foundation, Ministry of Education of Brazil, Braslia - DF 700040-020, Brazil.

## REFERENCES

Allen, B. F., Picon, F., Dalibard, S., Magnenat-Thalmann, N., and Thalmann, D. (2012). Localizing a mobile robot with intrinsic noise. In *3DTV-Conference: The True Vision - Capture, Transmission and Display of 3D Video (3DTV-CON), 2012*, pages 1–4.

Anderson, M. (2003). Embodied cognition: A field guide. *Artificial Intelligence*, 149(1):91–130.

Besag, J. (1986). On the Statistical Analysis of Dirty Pictures. *Journal of the Royal Statistical Society. Series B (Methodological)*, 48(3):259–302.

Bodiroza, S., Schillaci, G., and Hafner, V. (2011). Robot ego-sphere: An approach for saliency detection and attention manipulation in humanoid robots for intuitive interaction. In *2011 11th IEEE-RAS International Conference on Humanoid Robots (Humanoids)*, pages 689–694.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *JOURNAL OF THE ROYAL STATISTICAL SOCIETY, SERIES B*, 39(1):1–38.

Deng, H. and Clausi, D. (2004). Unsupervised image segmentation using a simple MRF model with a new implementation scheme. In *Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004*, volume 2, pages 691–694 Vol.2.

Dune, C., Herdt, A., Stasse, O., Wieber, P. B., Yokoi, K., and Yoshida, E. (2010). Cancelling the sway motion of dynamic walking in visual servoing. In *2010 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 3175–3180.

Exner, D., Bruns, E., Kurz, D., Grundhofer, A., and Bimber, O. (2010). Fast and robust CAMShift tracking. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 9–16.

Hoffmann, M. and Pfeifer, R. (2012). The implications of embodiment for behavior and cognition: animal and robotic case studies. *CoRR*, abs/1202.0440.

Hornung, A., Wurm, K., and Bennewitz, M. (2010). Humanoid robot localization in complex indoor environments. In *Intelligent Robots and Systems (IROS), 2010 IEEE/RSJ International Conference on*, pages 1690–1695.

Kato, Z., Pong, T.-C., and Chung-Mong Lee, J. (2001). Color image segmentation and parameter estimation in a markovian framework. *Pattern Recognition Letters*, 22(34):309–321.

Legrand, L., Bordier, C., Lalande, A., Walker, P., Brunotte, F., and Quantin, C. (2002). Magnetic resonance image segmentation and heart motion tracking with an active mesh based system. In *Computers in Cardiology, 2002*, pages 177–180.

Lewis M.A. and Simo L.S. (1999). Elegant stepping: A model of visually triggered gait adaptation. *Connection Science*, 11(3-4):331–344.

Lucas, B. D. and Kanade, T. (1981). An iterative image registration technique with an application to stereo vision. In *Proceedings of the 7th International Joint Conference on Artificial Intelligence - Volume 2*, IJCAI'81, pages 674–679, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics*, pages 281–297, Berkeley, Calif. University of California Press.

Miall, R. C., Weir, D. J., Wolpert, D. M., and Stein, J. F. (1993). Is the cerebellum a smith predictor? *Journal of motor behavior*, 25(3):203–216. PMID: 12581990.

Michel, P., Chestnutt, J., Kuffner, J., and Kanade, T. (2005). Vision-guided humanoid footstep planning for dynamic environments. In *Proceedings of the IEEE-RAS Conference on Humanoid Robots (Humanoids'05)*, pages 13 – 18.

Moughlbay, A., Cervera, E., and Martinet, P. (2013). Model based visual servoing tasks with an autonomous humanoid robot. In Lee, S., Yoon, K.-J., and Lee, J., editors, *Frontiers of Intelligent Autonomous Systems*, volume 466 of *Studies in Computational Intelligence*, pages 149–162. Springer Berlin Heidelberg.

Pretto, A., Menegatti, E., Bennewitz, M., Burgard, W., and Pagello, E. (2009). A visual odometry framework robust to motion blur. In *Robotics and Automation, 2009. ICRA '09. IEEE International Conference on*, pages 2250–2257.

Robert Cupec, G. S. (2005). Vision-guided walking in a structured indoor scenario. *Automatika*, 46(12):49–57.

Shapiro, L. (2007). The embodied cognition research programme. *Philosophy Compass*, 2(2):338–346.