Spatiotemporal Context in Robot Vision: Detection of Static Objects in the RoboCup Four Legged League

Pablo Guerrero, Javier Ruiz-del-Solar and Rodrigo Palma-Amestoy

Department of Electrical Engineering, Universidad de Chile

Abstract. Having as a main motivation the development of robust and high performing robot vision systems that can operate in dynamic environments, we propose a context-based generic vision system for a mobile robot with a mobile camera. We choose as a first application for this vision system, the detection of static objects in the RoboCup Four Legged League domain. Preliminary results using real video sequences are presented.

1 Introduction

Object visual perception in complex and dynamical scenes with cluttered backgrounds is a very difficult task, which humans can solve satisfactorily. However, computer and robot vision systems perform very badly in this kind of environments. One of the reasons of this large difference in performance is the use of context or contextual information by humans. Several studies in human perception have shown that the human visual system makes extensive use of the strong relationships between objects and their environment for facilitating the object detection and perception ([1][3][5][6][12], just to name a few).

Context can play a useful role in visual perception in at least three forms: (i) Reducing the perceptual aliasing: 3D objects are projected onto a 2D sensor, and therefore in many cases there is an ambiguity in the object identity. Information about the object surround can be used for reducing or eliminating this ambiguity; (ii) Increasing the perceptual abilities in hard conditions: Context can facilitate the perception when the local intrinsic information about the object structure, as for example the image resolution, is not sufficient; (iii) Speeding up the perceptions: Contextual information can speed up the object discrimination by cutting down the number of object categories, scales and poses that need to be considered.

From the visual perception point of view, it is possible to define at least six different types of context:

(i) Low-level context: Textures and color are perceived uniformly in objects parts and surfaces, independently of the illumination conditions and the presence of shadows or highlights. This is achieved using spatial diffusion mechanisms that interpolated low-level perceptions (interpolation at the pixel level). In humans this is carried out using cell mechanisms present in cortical areas V1-V4 [9].

(ii) Physical spatial context: There are physical laws that determine the allowed positions of the physical objects in the world. Once the observer knows in which

Guerrero P., Ruiz-del-Solar J. and Palma-Amestoy R. (2007).
Spatiotemporal Context in Robot Vision: Detection of Static Objects in the RoboCup Four Legged League.
In *Robot Vision*, pages 136-148
DOI: 10.5220/0002069901360148

physical context it is involved, it is possible to apply a corresponding visual model. A very general starting point for building a visual model is the assumption of the existence of a ground plane and a gravity vector, which allows us to define upward and downward directions. If we project the camera axis to the ground plane, then we can also define forward, backward, left and right. We can also define altitude as the distance to the ground plane. It is also possible to allow the existence of different horizontal planes in a single model, for example, if there is a table, over the ground, there can be other objects over the table. Most of the objects -more precisely, non-flying objects- either are supported by a horizontal plane or accelerate with gravity. Supported objects have an almost constant altitude, and their vertical orientation is usually almost constant and sometimes predetermined.

(iii) Temporal context: The cinematic models for the object's and observer's movements define their relative positions in different time steps. Thus, if an object is detected in a given position at time step k, then it should appear at a certain position in time step k+1.

(iv) Objects' configuration context: Normally physical objects are seen in specific spatial configurations or groups. For instance, a computer monitor is normally observed near a keyboard and a mouse; or a face, when detected in its normal upright pose, it is seen above the shoulders and below hair.

(v) Scene context: In some specific cases, scenes captured in images can be classified in some defined types [8], as for examples "sunset", "forest", "office environment", "portrait", etc. This scene context, which can be determined using a holistic measurement from the image [1][2][7] and/or the objects detected in the same image, can contribute to the final detection or recognition of the image's objects.

(vi) Situation context: A situation is defined by the surround in which the observer is immersed (environment and place), as well as by the task being performed. An example of a situation context could be: "playing tennis in a red clay curt, in a sunny day, at 3PM". The situation context is determined using several consecutive visual perception, as well as other source of perceptual information (e.g. auditory) and high-level information (e.g. task being carried out, weather, time of the day).

In [12] are also defined the photometric context (the information surrounding the image acquisition process, mainly intrinsic and extrinsic camera parameters), and also the computational context (the internal state of processing of the observer). However, we believe that those do not correspond to contextual information in the same sense we are defining it in this work.

Low-level context is frequently used in computer vision. Thus, most of the systems performing color or texture perception uses low-level context to some degree (see for example [13]). Scene context have been also addressed in some computer vision [10] and image retrieval [4] systems. However, we believe that not enough attention has been given in robot and computer vision to the physical-spatial context, the temporal context, the objects' configuration context, and the situation context.

Having as our main motivation the development of robust and high performing robot vision systems that can operate in dynamic environment in real-time, in this work we propose a generic vision system for a mobile robot with a mobile camera, which employs all defined spatiotemporal contexts. We strongly believe that as in the case of the human vision, contextual information is a key factor for achieving high performance in dynamic environments. Although other systems, as for example [1][3]

[5][12] have also employed contextual information, to the best of our knowledge this is the first work in which context is addressed in an integral fashion. We believe that the use of a context filter that takes into account the coherence between current and past detections, as well as scene and situation contexts, and the estimation of the position of static objects when they are out of the field of view, are some of the most innovative contributions of this work.

We choose as a first application for our vision system, the detection of static objects in the RoboCup Four Legged League domain. We select this application domain mainly because static objects in the field (beacons, goals and field lines) are part of a fixed and previously known 3D layout, where it is possible to use several relationships between objects for calculating the defined contexts.

This paper is organized as follows. The proposed generic vision system for a mobile robot with a mobile camera is described in section 2. In section 3, this general vision system is adapted for the detection of static objects in the RoboCup 4L League. Finally, conclusions of this work are given in section 4.

2 Proposed Context based Vision System

The proposed vision system is summarized in the block diagram shown in figure 1. It considers object preceptors, a holistic characterization of the scenes, context coherence filtering between current and past detections, encoder-based, visual-based and filtered horizon information, objects characterization, situation (global) context, and high-level tracking (estimation) for the detected objects' poses.

2.1 Perceptors

Each object of interest has a specialized perceptor that detects and characterizes it. For a detailed description of the perceptors used in our system, refer to [16]. The output of a perceptor *i* at time step *k* is a *candidate object* c_k^i, defined as:

$$c_k^i = \left(\mathbf{x}_k^i, \Sigma_{\mathbf{x}_k^i}, \alpha_k^i, \mathbf{y}_k^i, \eta_k^i, \sigma_{\mathbf{y}_k^i}, \sigma_{\eta_k^i} \right)^T \tag{1}$$

where $\mathbf{x}_k^i$ is the relative pose of the object with respect to the robot, $\Sigma_{\mathbf{x}_k^i}$ is the covariance matrix of $\mathbf{x}_k^i$, α_k^i is the a priori probability of the detection, and ($\mathbf{y}_k^i$, η_k^i) and ($\sigma_{\mathbf{y}_k^i}$, $\sigma_{\eta_k^i}$) are the horizon position and angle, and their corresponding tolerances (see explanation in section 2.5.2).

2.2 Image Holistic Characterization

As stated in [6], a single glance to a complex, real world scene is enough for an observer to comprehend a variety of perceptual and semantic information. There are several works that use different alternatives of representations of the global

information contained in an image (e.g. spatial frequency orientations and scales, color, texture density). We believe that some of those representations can be complementary, and that the selection of a subset of them can be done taking into account the kind of application for which the vision system is designed. For a general mobile robot vision system it is expected that the image characterization be invariant to lighting conditions, rotations, translations, and image distortions such as blur.

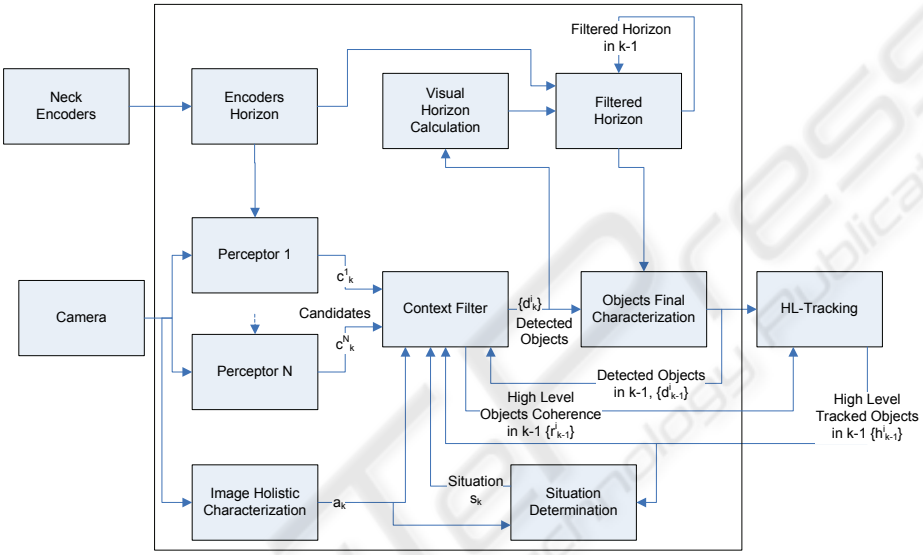


Fig. 1. Block diagram of the proposed general vision system for a mobile robot with a mobile camera.

2.3 Situation Determination

This stage intends to determine the situation in which the observer is involved. We propose that, for solving this task, it is necessary to consider several consecutive images, in an incremental process. This is inspired in the hypothesis that human visual system needs a short-term memory to adequately interpret images [11]. What we call as a situation goes from the kind of environment the robot is immersed in (natural, artificial), passing through the determination of the specific place where the robot is (inside room X, inside building Y), to the specific task that the robot is carrying out (crossing the street, raising the stairs, in a soccer game, etc.). A situation is not expected to change suddenly. Instead, one expects to have transition probabilities between situations given the observations and the observer odometry. We think that the situation state can be modeled as a Markov process, where, for each known situation, there is a transition probability to other situation. The transition probabilities are conditioned on the last coherent observations (HL-tracking module), and the current image holistic characterization.

2.4 H-L Tracking

The high-level (H-L) tracking stage is intended to maintain information about the objects detected in the past, although they are currently not observed (for instance, in any moment you have an estimation of the relative position of the objects that are behind you). This tracking stage is basically a state estimator for each object of interest, where the state to be estimated, for fixed objects, is the relative pose $\mathbf{x}_k^i$ of the object with respect to the robot. For mobile objects, the relative velocity $\mathbf{v}_k^i$ may be added to the state vector. Every time the robot moves, a new state is predicted for each object been tracked. On the other hand, every time an object is detected, its state estimation $\mathbf{x}_k^i$ and covariance $\Sigma_{\mathbf{x}_k^i}$ are corrected based in the new observation. This module can be implemented using standard state estimation algorithms as Kalman or Particle Filters.

2.5 Context Filter

In this module context information is employed for filtering candidate objects. All candidate objects must be coherent between them, and also they must be coherent with the current situation, and the holistic image characterization. Moreover, candidate objects must be coherent with their past detections (detected objects in the previous time step), and their current pose estimation. Thus, the context of a candidate object consists of all the information obtained from the current and past images, excepting the information obtained from the candidate itself.

2.5.1 Objects Coherence

Let $\mathbf{K}_k^i$ be the vector of N candidates from the image at time k, excepting c_k^i, $\mathbf{D}_k$ the vector of N detections obtained by applying the context coherence filter to the elements of $\mathbf{C}_k$, $\mathbf{H}_k$ the vector of high-level objects poses estimations, a_k the current image holistic characterization, and s_k the current situation. Then, the context vector $\mathbb{C}_k^i$ of c_k^i is defined as:

$$\mathbb{C}_k^i = \left(\mathbf{K}_k^{i\,T}, \mathbf{D}_{k-1}{}^T, \mathbf{H}_{k-1}{}^T, a_k, s_k\right)^T \tag{2}$$

Every element $\left[\mathbb{C}_k^i\right]_j$ of the context vector has an associated weight $\omega_{k,j}^i$ that corresponds to the probability of $\left[\mathbb{C}_k^i\right]_j$. The context weight vector is defined as:

$$\mathbf{\Omega}_k^i = \left(\mathbf{\Omega}_k^{\mathbf{C},iT}, \mathbf{\Omega}_{k-1}^{\mathbf{D}\,T}, \mathbf{\Omega}_{k-1}^{\mathbf{H}\,T}, p(a_k), p(s_k)\right)^T = \left(\omega_{k,1}^i, ..., \omega_{k,L}^i\right)^T \tag{3}$$

with $L = 3N + 1$, and

$$\mathbf{\Omega}_k^{\mathbf{C},i} = \left(\alpha_k^1,..,\alpha_k^{i-1},\alpha_k^{i+1},..,\alpha_k^N\right)^T;\ \mathbf{\Omega}_{k-1}^{\mathbf{p}} = \left(p_{k-1}^1,..,p_{k-1}^N\right)^T;\ \mathbf{\Omega}_{k-1}^{\mathbf{H}} = \left(\beta_{k-1}^1,..,\beta_{k-1}^N\right)^T \quad (4)$$

Here, α_k^i, p_k^i and β_k^i are respectively the a priori probability of the candidate c_k^i, the a posteriori probability of the detection d_k^i, and the accumulated probability of the tracked object h_k^i. Then, we define the coherence of c_k^i as:

$$q_k^i = p\left(c_k^i \mid \mathbb{C}_k^i\right) = \frac{\sum_{j=1}^{L} p\left(c_k^i \mid \left[\mathbb{C}_k^i\right]_j\right) p\left(\left[\mathbb{C}_k^i\right]_j\right)}{\sum_{j=1}^{L} p\left(\left[\mathbb{C}_k^i\right]_j\right)} = \frac{\sum_{j=1}^{L} p\left(c_k^i \mid \left[\mathbb{C}_k^i\right]_j\right)\omega_{k,j}^i}{\sum_{j=1}^{L}\omega_{k,j}^i} \quad (5)$$

The a posteriori probability of c_k^i is then defined as:

$$p_k^i = \alpha_k^i q_k^i = \alpha_k^i \frac{\sum_{j=1}^{L} p\left(c_k^i \mid \left[\mathbb{C}_k^i\right]_j\right)\omega_{k,j}^i}{\sum_{j=1}^{L}\omega_{k,j}^i} \quad (6)$$

We shall then calculate the probabilities $p\left(c_k^i \mid \left[\mathbb{C}_k^i\right]_j\right)$.

2.5.2 Relationships between Physical Objects

There are four kind of relationships that can be checked between physical objects. The first two must be checked between candidates belonging to the same image, or at most between candidates of very close images, when the camera's pose change is bounded. The last two may be checked between candidates or objects of different images, considering the camera's pose change between images.

Horizontal Position Alignment. As we said before, most of the objects in the real world are supported on ground. This, plus the knowledge of the height of the object, gives us information about a region of the object that is likely to be at the same altitude as the camera. For each candidate c_k^i, we will call the center of this region, $\mathbf{y}_k^i$, the *horizontal point* of the candidate, and its tolerance will be noted $\sigma_{\mathbf{y}_k^i}$. Horizontal points of correct candidates are supposed to be part of a line, the visual horizon.

Horizon Orientation Alignment. Another quality of several objects is their almost fixed orientation with respect to a vertical axis. Using this quality, it is possible to find a horizon angle that is coherent with the orientation of the object in the image. For each candidate c_k^i, we will call this angle, η_k^i, the horizontal angle of the candidate, and its tolerance will be noted $\sigma_{\eta_k^i}$. Horizontal angles of correct candidates must have similar values, and furthermore, they are expected to be similar to the angle of the visual horizontal obtained from the horizontal points.

Relative Position or Distance Limits. In some specific situations, objects are part of a fixed layout. The robot may know this layout a priori from two different sources: been previously taught about it, or learning it from observations, as in SLAM. In both cases, the robot can check if the relative position between two objects, or at least their distances (when objects has radial symmetry), is maintained.

Speed and Acceleration Limits. Even between mobile objects, it is possible to check their relative speed and acceleration limits. Of course, it is necessary to have a previous knowledge of those limits related with the objects identity.

2.5.3 High Level Tracked Objects Maintenance

When an object is detected and it is not been tracked, the HL-Tracking module creates a new state estimator for it, and initializes it whit all the values coming from the detection process. In particular, the coherence is initialized with the coherence obtained by the candidate that generated the detection. However, as the robot moves, odometry errors accumulate and high-level estimations become unreliable. If a set of high-level estimations is self-coherent, but moves too far from real poses of tracked objects, then all the new observations may become incoherent and be rejected. To avoid this kind of situations, high-level estimations are also evaluated in the coherence filter. In order to inhibit the self-confirmation of an obsolete set of estimations, the coherence r_k^i is only checked with respect to the current observations, but it is smoothed to avoid a single outlier observation discarding all the objects been tracked. Thus, the coherence of a tracked object is updated using:

$$r_k^i = \lambda r_{k-1}^i + (1-\lambda)\frac{\sum_{j=1}^{N} p\left(h_k^i \mid c_k^j\right)\alpha_k^j}{\sum_{j=1}^{N}\alpha_k^j} \tag{7}$$

where λ is a smoothing factor. As the coherence is recalculated, the a posteriori score shall be also recalculated according to $\beta_k^i = r_k^i \alpha_k^i$. The a priori score is kept in the same value obtained from the object's last perception.

2.6 Encoders Horizon, Visual Horizon, Filtered Horizon and Objects Final Characterization

In the *Encoders Horizon* module, the horizon line in each image is estimated using the information from the neck encoders, and an estimation of the body inclination obtained from either encoders or accelerometers. However, depending on the robot architecture, both measurements used to calculate the encoders' horizon are noisy and can yield a very poor estimate of the horizon line. This affects the characterization of some objects, for example, when measuring the distance to certain pixel which is known to be at the ground, using its elevation angle. To solve this problem, we have implemented a *Visual Horizon* module. In this module, information from detected objects is used to calculate a consensual horizon. The horizon line can be defined with

two parameters: its distance to the image center and its angle with respect to the rows direction of the image.

Even when no visual object is detected, the previously computed horizon and the changes of neck-encoders' measurements are used to predict a final horizon in the *Filtered Horizon* module. Afterwards, in the *Objects Final Characterization* module objects relative poses are recalculated using the filtered horizon.

3 Detection of Static Objects in the RoboCup 4L League

We apply the proposed vision system in the RoboCup Four Legged League environment (see description in [14]). In a Four Legged League soccer field, there are many objects that have spatial relationships between them, and defined temporal behaviors. These objects are goals, beacons, a ball, robots, and field lines. Some of the objects are static and others are moving. Goals, beacons and field lines detection is essential for the robot self-localization in the field, while ball and robots detection is necessary for correctly playing soccer. Static objects in the field (beacons, goals and field lines) are part of a fixed and previously known 3D layout. Thus, it is possible to use several of the proposed relationships between objects to calculate a candidate's coherence. For this reasons we choose as a first application of our vision system, the detection of static objects in the RoboCup Four Legged League domain.

Beacons have a radial symmetry. Detecting goals rotation with respect to the camera is difficult because occlusions may change the ratio between the goal's height and width. For that reason it is not possible to determine the relative position of one of these objects relative to other. Nevertheless we are able to use distances between them and laterality. Laterality information comes from the fact that the robot is always moving in an area that is surrounded by the fixed objects. For that reason, it is always possible to determine, for any pair of candidates, which of them should be to the right of the other. Then, depending on the kind of information that $\left[\mathbb{C}_k^i\right]_j$ represents, $p\left(c_k^i \mid \left[\mathbb{C}_k^i\right]_j\right)$ is calculated as:

$$p\left(c_k^i \mid c_k^j\right) = p_{Hor}\left(c_k^i \mid c_k^j\right) p_{Lat}\left(c_k^i \mid c_k^j\right) p_{Dist}\left(c_k^i \mid c_k^j\right) \tag{8}$$

$$p\left(c_k^i \mid d_{k-1}^j\right) = p_{Hor}\left(c_k^i \mid d_{k-1}^j\right) p_{Lat}\left(c_k^i \mid d_{k-1}^j\right) p_{Dist}\left(c_k^i \mid d_{k-1}^j\right) \tag{9}$$

$$p\left(c_k^i \mid h_{k-1}^j\right) = p_{Lat}\left(c_k^i \mid h_{k-1}^j\right) p_{Dist}\left(c_k^i \mid h_{k-1}^j\right) \tag{10}$$

$$p\left(c_k^i \mid a_k\right) = p\left(\mathbf{y}_k^i, \eta_k^i \mid a_k\right) \tag{11}$$

We have not yet implemented the use of s_k, as our vision system has been tested in one single application.

The horizontal coherence between two candidates is approximated using a triangular function:

$$p_{Hor}\left(c_k^i \mid c_k^j\right) = tri\left(\Delta\eta_k^{i,j}, \sigma_{\Delta\eta_k^{i,j}}\right) tri\left(\Delta\eta_k^{j,i}, \sigma_{\Delta\eta_k^{j,i}}\right) \tag{12}$$

$$tri\left(\Delta x, \sigma\right) = \begin{cases} 1 - \dfrac{\Delta x}{\sigma} & \Delta x < \sigma \\ 0 & otherwise \end{cases} \tag{13}$$

with

$$\Delta\eta_k^{i,j} = \left|\eta_k^i - \eta_k^{i,j}\right| \;;\; \eta_k^{i,j} = \eta_k^{j,i} = \measuredangle\left(\mathbf{y}_k^i - \mathbf{y}_k^j\right) \tag{14}$$

and

$$\sigma_{\Delta\eta_k^{i,j}} = \sigma_{\Delta\eta_k^{j,i}} = \sigma_{\eta_k^i} + \sigma_{\eta_k^j} + \tan^{-1}\left(\frac{\sigma_{\mathbf{y}_k^i} + \sigma_{\mathbf{y}_k^j}}{\left|\mathbf{y}_k^i - \mathbf{y}_k^j\right|}\right) \tag{15}$$

Similarly, the calculation of $p_{Hor}\left(c_k^i \mid d_k^j\right)$ is totally analogous with only two differences: $\mathbf{y}_k^j$ and η_k^j are modified using the encoders information, and the tolerances $\sigma_{\eta_k^j}$ and $\sigma_{\mathbf{y}_k^j}$ are increased to meet the uncertainty generated by the possible camera and robot movements. The lateral coherences, $p_{Lat}\left(c_k^i \mid c_k^j\right)$, $p_{Lat}\left(c_k^i \mid d_k^j\right)$ and $p_{Lat}\left(c_k^i \mid h_k^j\right)$, are defined as binary functions, which are equal to 1 if the lateral relation between c_k^i and c_k^j, d_k^j or h_k^j, is the expected one, and 0 otherwise. The distance coherence, p_{Dist}, is also approximated using a triangular function:

$$p_{Dist}\left(c_k^i \mid c_k^j\right) = tri\left(\Delta\mathbf{x}_k^{i,j}, \sigma_{\Delta\mathbf{x}_k^{i,j}}\right) \tag{16}$$

with $\Delta\mathbf{x}_k^{i,j} = \left|\mathbf{x}_k^i - \mathbf{x}_k^j\right|$, and $\mathbf{x}_k^i$, $\mathbf{x}_k^j$ been the relative detected positions of c_k^i and c_k^j respectively. The calculation of $p_{Dist}\left(c_k^i \mid d_k^j\right)$ and $p_{Dist}\left(c_k^i \mid h_k^j\right)$ is analogous. The tolerance $\sigma_{\Delta\mathbf{x}_k^{i,j}}$ is calculated as a function of the covariance matrices $\Sigma_{\mathbf{x}_k^i}$ and $\Sigma_{\mathbf{x}_k^j}$.

The global measure of the image we selected for this application is similar to that of Color Shapes [15], with the difference that we make histograms of color classes, after a color segmentation stage. We divided the image in cells of 16x16 pixels. The probability $p\left(\mathbf{y}_k^i, \eta_k^i \mid a_k\right)$ is approximated by:

$$p\left(\mathbf{y}_k^i, \eta_k^i \mid a_k\right) = p\left(a_{\inf,k}^i\right) p\left(a_{\sup,k}^i\right) \tag{17}$$

where $a_{\inf,k}^i$ and $a_{\sup,k}^i$ are respectively the inferior and superior histograms of the image. The superior (inferior) histogram counts all bins of cells being over (below) the horizon line. The probabilities $p\left(a_{\inf,k}^i\right)$ and $p\left(a_{\sup,k}^i\right)$ are associated to the

likelihoods of $a^i_{\text{inf},k}$ and $a^i_{\text{sup},k}$ been obtained from the corresponding regions of the image:

$$p\left(a^i_{\text{inf},k}\right) = 1 - \frac{innov\left(a^i_{\text{inf},k}, \overline{a}_{\text{inf}}, \mathbf{A}_{\text{inf}}\right)}{innov\left(a^i_{\text{inf},k}, \overline{a}_{\text{inf}}, \mathbf{A}_{\text{inf}}\right) + innov\left(a^i_{\text{inf},k}, \overline{a}, \mathbf{A}\right)} \tag{18}$$

$$p\left(a^i_{\text{sup},k}\right) = 1 - \frac{innov\left(a^i_{\text{sup},k}, \overline{a}_{\text{sup}}, \mathbf{A}_{\text{sup}}\right)}{innov\left(a^i_{\text{sup},k}, \overline{a}_{\text{sup}}, \mathbf{A}_{\text{sup}}\right) + innov\left(a^i_{\text{sup},k}, \overline{a}, \mathbf{A}\right)} \tag{19}$$

with $innov\left(\mathbf{z}, \overline{\mathbf{z}}, \mathbf{Z}\right) = \left(\mathbf{z} - \overline{\mathbf{z}}\right)^T \mathbf{Z}\left(\mathbf{z} - \overline{\mathbf{z}}\right)$.

This has the implicit assumption that $a^i_{\text{inf},k}$ and $a^i_{\text{sup},k}$ has different PDF's, which in our application is valid (for example, it is more likely to find green below the horizon). The parameters $\overline{a}_{\text{sup}}, \mathbf{A}_{\text{sup}}, \overline{a}_{\text{inf}}, \mathbf{A}_{\text{inf}}, \overline{a}, \mathbf{A}$ correspond to the means and covariances over and below the horizon and of the whole image. They are obtained from a training set of approximately 200 images captured from different positions in the field. $p\left(a_k\right)$ is assumed to be Gaussian with mean $\overline{a}$ and covariance $\mathbf{A}$.

4 Experimental Results

We have tested our vision system using real video sequences obtained by an AIBO Robot inside a RoboCup four legged soccer field. The detection rates were measured in different situations having different quantities of false objects. False objects are objects that resemble to the actual ones, and that the color-based vision system is not able to correctly distinguish. For example, a spectator wearing a yellow shirt is a false yellow goal. Even objects having the same appearance but put in a different place are false objects. For example, a yellow goal of a neighbor field is a false yellow goal.

For illustrating the operation of the system, Fig. 2 shows some example images where context becomes relevant to discriminate between false and true objects. Fig. 2 shows two examples of a posteriori probabilities obtained by different false and true objects. Fig 2.a. contains a false beacon (p = 0.10), a false goal (p = 0.01), a true beacon (p = 0.38) and a true goal (p = 0.48). Fig 2.b. contains three beacons of which two are true (p = 0.43, 0.58) and one is false (p = 0.00). Even when the shown false objects are not a priori differentiable (they are all true objects placed in wrong places), and thus, a priori probabilities are the same for all objects (p = 1.00), context is helpful to differentiate a posteriori probabilities of true and false objects.

(a) (b)

Fig. 2. Some examples of a posteriori probabilities for true and false objects.

With the purpose of evaluating the performance of the system, the different ROC curves in Fig. 3 compare the use of the different scores of the object to detect it; a priori (with no use of context), a posteriori (using of all the context information), and partial scores obtained from the consideration of one single type of context: current (objects detected in the current image), HL-Tracking, last (objects detected in the last image), and global (global measurements). In the situation of Fig 3.a, false objects present were "natural" objects, like the cyan blinds and some other objects of our laboratory, these objects appear in approximately 20% of the frames. In Fig 3.b, two more aggressive objects were added: two false goals over the ground plane and in the border of the field. In Fig 3.c, two additional false objects were added (over those of Fig 3.b): two false beacons over the ground plane, in the border of the field.

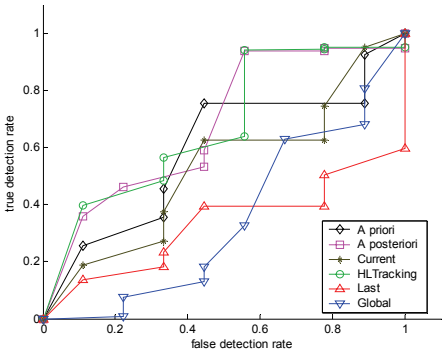
These noisy situations may appear as artificially overexposed to false objects, but they are neither very different to the actual situations observed in RoboCup games with several spectators in the border of the field nor more noisy than real environments situations. Note how the a priori and a posteriori ROC curves evolve as the quantity of noise is increased. When facing situations with a low quantity of false objects, the use of context appears as no improving the performance of the system (Fig 3.a). However, as the quantity of false objects grows, the use of context increases noticeably the detection rate for a given false negative rate (Fig 3.b, 3.c).

The results presented correspond to sequences obtained with the robot camera moving, and the robot either standing or moving.

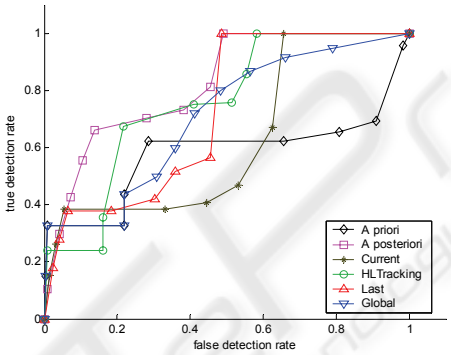
5 Conclusions

We have presented a general context based vision system for a mobile robot having a mobile camera. The use of spatiotemporal context is intended to make the vision system robust to noise and high performing in the task of object detection.

(a)



(b)



(c)

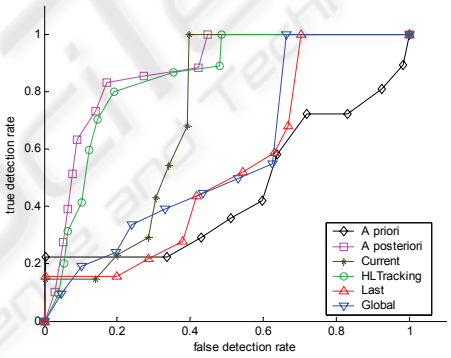


Fig. 3. ROC Curves with low (a), medium (b) and high (c) quantity of noise.

We have first applied our vision system to detect static objects in the RoboCup Four Legged League domain, and preliminary experimental results are presented.

Experimental results confirm that the use of spatiotemporal context is of great help to improve the performance obtained when facing the task of object detection in a noisy environment. We are working in obtaining more experimental results to provide a better characterization of the system. The existing results encourage us to continue developing our system and to test it in other applications, where different physical objects and lighting conditions may exist and thus a situation determination stage and different perceptors and global measures should be considered.

Acknowledgements

"This research was partially supported by FONDECYT (Chile) under Project Number 1061158".

References

1. A. Torralba, P. Sinha. "On Statistical Context Priming for Object Detection". *International Conference on Computer Vision*, 2001.
2. A. Torralba. "Modeling global scene factors in attention". *JOSA* - A, vol. 20, 7, 2003.
3. D. Cameron and N. Barnes. "Knowledge-based autonomous dynamic color calibration". *The Seventh International RoboCup Symposium*, 2003.
4. A. Oliva, A. Torralba, A. Guerin-Dugue, and J. Herault. "Global semantic classification of scenes using power spectrum templates". *Proceedings of The Challenge of Image Retrieval (CIR99)*, Springer Verlag BCS Electronic Workshops in Computing series, Newcastle, UK., 1999.
5. M. Jüngel, J. Hoffmann and M. Lötzsch. "A real time auto adjusting vision system for robotic soccer". *The Seventh International RoboCup Symposium*, 2003.
6. A. Oliva. "Gist of the Scene". *Neurobiology of Attention*. Elsevier, San Diego, CA, pp. 251-256. 2003.
7. S. Foucher, V. Gouaillier and L. Gagnon. "Global semantic classification of scenes using ridgelet transform". *Human Vision and Electronic Imaging IX. Proceedings of the SPIE*, Volume 5292, pp. 402-413. 2004.
8. A. Torralba and A. Oliva, "Statistics of Natural Image Categories". *Network: Computation in Neural Systems*, No 14, August, pp. 391-412, 2003.
9. L. Spillman and J. Werner (Eds.), *Visual Perception: The Neurophysiological Foundations*, Academic Press, 1990.
10. A. Oliva, and A. Torralba. "Modeling the Shape of the Scene: A Holistic Representation of the Spatial Envelope". *International Journal of Computer Vision*, Vol. 42, No. 3, pp. 145-175. 2001.
11. Potter, M. C., Staub, A., Rado, J., & O'Connor, D. H. "Recognition memory for briefly presented pictures: The time course of rapid forgetting". *Journal of Experimental Psychology. Human Perception and Performance*, 28, pp. 1163–1175. 2002.
12. Strat, T. "Employing contextual information in computer vision". *Proceedings of DARPA Image Understanding Workshop*. 1993.
13. J. Ruiz-del-Solar and R. Verschae, "Skin Detection using Neighborhood Information". *Proc. 6th Int. Conf. on Face and Gesture Recognition – FG 2004*, 463 – 468, Seoul, Korea, May 2004.
14. RoboCup Technical Comitee, "RoboCup Four-Legged League Rule Book". http://www.tzi.de/4legged/bin/view/Website/WebHome. 2006.
15. R. Stehling, M. Nascimento, and A. Falcao. "On 'Shapes' of Colors for Content-Based Image Retrieval". *Proceedings of the International Workshop on Multimedia Information Retrieval*, pp 171-174. 2000.
16. Zagal, J.C., Ruiz-del-Solar, J., Guerrero, P. and Palma R. (2004). "Evolving Visual Object Recognition for Legged Robots". Lecture Notes in Computer Science 3020 (RoboCup 2003), Springer, 181-191.