Combining local and global visual information in context-based neurorobotic navigation

Marwen Belkaid, Nicolas Cuperlier, Philippe Gaussier

Combining local and global visual information in context-based neurorobotic navigation

Marwen Belkaid Nicolas Cuperlier Philippe Gaussier
Neurocybernetics team, ETIS laboratory, UMR 8051, CNRS/ENSEA/UCP
95000 Cergy Cedex, France.
{*firstname.lastname*}@ensea.fr

***Abstract*— In robotic navigation, biologically inspired localization models have often exhibited interesting features and proven to be competitive with other solutions in terms of adaptability and performance. In general, place recognition systems rely on global or local visual descriptors; or both. In this paper, we propose a model of context-based place cells combining these two information. Global visual features are extracted to represent visual contexts. Based on the idea of global precedence, contexts drive a more refined recognition level which has local visual descriptors as an input. We evaluate this model on a robotic navigation dataset that we recorded in the outdoors. Thus, our contribution is twofold: 1) a bio-inspired model of context-based place recognition using neural networks; and 2) an evaluation assessing its suitability for applications on real robot by comparing it to 4 other architectures – 2 variants of the model and 2 stacking-based solutions – in terms of performance and computational cost. The context-based model gets the highest score based on the three metrics we consider – or is second to one of its variants. Moreover, a key feature makes the computational cost constant over time while it increases with the other methods. These promising results suggest that this model should be a good candidate for a robust place recognition in wide environments.**

I. Introduction

Place recognition is one of the abilities that mobile robots need in order to navigate properly. In this context, taking inpiration from biological systems has often been a good way to build robust and adaptive systems that can also be competitive with non-biologically inspired models in terms of performance [1][2][3]. Indeed, mammals have the capacity to recognize visited places and to localize themselves in nearly any sort of environments. Place cells, grid cells and head direction cells represent the neural substrate involved in this capacity and have been extensively studied in the last decades (see [4] for a recent review of navigation computational models). Generally, vision is the most important modality – although other sensory inputs (e.g. olfactory, tactile, proprioceptive, etc.) are also very helpful.

In the literature, we can classify scene recognition methods according to the type of visual information they rely on, i.e. global or local. Some of them try to capture the global aspect of the scene by considering the image as a whole [5][2][6][3][7]. On the other hand, other models extract and encode particular subparts of the image, called regions of interest [8][1][9][10]. While global vision-based methods are faster, they have generally shown less effective than local vision-based ones [11][3].

The issue of combining local and global visual information has been addressed in related work [12][3][13][6]. In this paper, we propose a biologically inspired model for context-based place recognition. It is based on two parallel pathways learning to discriminate the robot locations from global low-resolution and local high-resolution visual inputs respectively. The first neural network serves for the contextualization of the information being processing in the second one. This hierarchical model explores the idea of visual contexts [14][15] in the case of global precedence [16][5].

The hierarchical combination of local and global information represented by our model is compared to a stacking approach (concatenation of inputs). We also use a model of place recognition that is purely local vision-based as a baseline for this comparison. Those architectures are tested on a dataset recorded in visually different outdoor environments. Also, it fits the protocol used in experiments on real robots [17]. Moreover, we introduce a set of measures that not only evaluate the vision system from the information retrieval standpoint but also in the specific case of robotic navigation. Thereby, we show that our context-based approach outperforms the other methods and is viable in terms of computational cost. Our results suggest a good potential for scaling up to wide environments.

In the next section, we elaborate on the differences between local and global vision in order to introduce the notion of visual contexts in the case of a localization task. Then, we introduce the model of context-based place recognition using neural networks. Next, we describe our dataset and give some implementation details regarding the visual descriptors. Finally, two experiments are presented. The first one compares the descriptors capacity to discriminate places in terms of granularity. The second one compares our model to 4 other architectures – 2 variants of the model and 2 stacking-based solutions – in order to evaluate their performance and computational cost.

II. Local Vs Global Vision

Scene recognition models can be classified based on whether they use local or global visual descriptors. Global (holistic) vision methods consider the image as a whole and

TABLE I
COMPARISON BETWEEN LOCAL RHOTHETA, SIFT AND SURF

	rhotheta	SIFT	SURF
Multiscale	No	Yes	Yes
Saliency map	DoG	DoG	DoB(Box)
PoI extraction	Local extrema	Local extrema	Hessian matrix
Descriptor type	Log polar mapping	Orientation histogram	Orientation histogram

encode it as a single vector. Because they are typically very compact, they allow for fast computation.

A significant number of models using this kind of descriptors can be found in the literature [5][2][6][3][7]. For instance, Milford proposes to subsample the whole image and use it as a global signature [6]. In former work [2], the subsampled image is also projected over the horizontal axis to obtain a vector representing the intensity profile. Global descriptors can be encoded in histograms as well [11][7]. In addition, some solutions construct signatures based on statistical moments (mean, variance or possibly higher order moments) [11][3]. Three types of information are commonly used to compute visual descriptors. First and foremost evident is the luminance channel [2][6]. Also, chrominance channels can provide richer information [11][7]. Lastly, orientations are very useful to describe textures both indoors (e.g. doors and computers screens) and outdoors (e.g. trees, roads and distant buildings) [5]. In [3], the authors use these three types of information simultaneously.

In contrast, local descriptors only carry information relative to certain regions of interest in the image. State-of-the-art methods, like SIFT [9] and SURF [10] typically implement this kind of solution. In previous work, we proposed a biologically plausible model for place cells driven by visual input obtained from an attentional system [8][18]. Also, studies using real robots showed its robustness in indoor and outdoor environments [1][17]. In this model, points of interest are extracted from the image based on a saliency map. Local views around these salient points are transformed using log-polar mapping and used as descriptors. The Table I summarizes the differences between this technique (labelled *rhotheta*) and SIFT [9] and SURF [10] techniques. In our place cells model, the *rhotheta* codes (i.e. local views encoded using log-polar transform) are categorized and used to build a representation of a visited place by merging "what" and "where" information in a 2D map using a max-pi operation (max of tensor product). We will refer to this whole method as LPMP (Log-Polar Max-Pi). More details will be given in Sect. IV.

Generally, holistic methods have shown less effective than local features-based ones [11][3]. Indeed, the latter use richer information: a certain number of signatures per image ("what" information) and their corresponding positions ("where" information). However, they can be useful in situations where stability of points of interest detection is difficult to ensure (complex textures like tree leaves, condition variation, etc.). Prior to the experiments we present in this paper, we conducted a pilot study that allowed us to evaluate the descriptors separately and set the values of the parameters we list later on. We cannot show the results here due to space limitation. However, the overall conclusion is that, although the local descriptors generally get higher scores, holistic descriptors perform well and sometimes outperform the local ones.

Some models combine local and global descriptors to benefit from both kinds of visual features [12][3][13][6]. For example, the two levels of visual descriptors can be concatenated and fed simultaneously to the scene classifier [12][13]. In other cases, the global features are used for a first level of recognition (i.e. bigger regions of the navigation environment), which is then refined by the local information (i.e. more precise location). We refer to the first category as *stacking* methods and the second one as *hierarchical* methods [12]. The model we present in the paper belongs to the second class and explores the idea of visual contexts.

III. VISUAL CONTEXTS

Humans are able to coarsely recognize a scene at a glance. The *gist*[1] of the scene is extracted from its global aspect based on low-resolution visual information [16][15][14][5]. Such visual context identification resolves ambiguities and facilitates objects recognition – which is, in contrast, based on the processing of higher resolution information [14]. Indeed, contextual cueing has been shown to drive spatial attention and increase performances in search tasks [15].

The notion of context is very related to spatial representation as it often refers to background cues [19]. So context identification is essential for navigation. But, while place recognition *per se* is mainly based on geometric visual information (distances and directions of the landmarks in the environment), it is also influenced by additional sensory (e.g. colors, sounds, odors, etc.) and behavioral cues [20]. Thus, the term "context" can carry more or less abstract meanings in the literature [19].

In a pure localization tasks, context detection can refer to the recognition of a broad area in which several locations (places) can be discriminated. We previously proposed a model in which a coarse place recognition modulates the responses of a more refined recognition level [21]. This initial work provides proof of the interest of the contextualization of place recognition in a simulated navigation task. Yet, the two levels were based on the same geometric cues. On the other hand, according to the global precedence concept, the global aspect of the image can be used to drive the recognition of local details [16]. The pilot study we mentioned earlier gave preliminary evidence on real data that holistic descriptors can successfully discriminate large regions of the environment. For instance, Fig. 1 shows that two of the global descriptors we use in this paper are able to discriminate the three datasets. The first experiment we present in this paper further

[1]The term "gist" is used in its general meaning – that is to say, the summary or the essential aspect of the scene. We do not refer to the particular implementation used in [3].

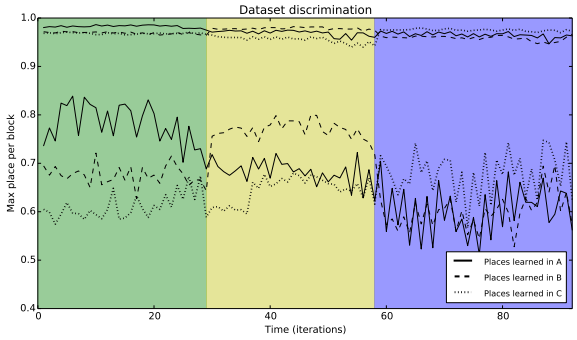

Fig. 1. Maximum activities among places that were respectively learned in A, B and C are represented (see Sect. V-A). Due to space limitation, we show the output of two experiments using *proflum* and *profcol* global descriptors on the same figure. The implementation of these two holistic descriptors is detailed in Sect. V-B.2. Despite differences in the dynamics, these descriptors correctly discriminate the different dataset.

shows the difference in the granularity of place recognition based on local or global visual information (see Sect. VI).

IV. Context-based Place Recognition Model

In this model, two identical neural networks are respectively dedicated to global low-resolution and local high-resolution visual processing (see Fig. 2). The former pathway modulates the activity of the latter. As a result, a coarse localization (*Context*) and a refined one (*Place*) are obtained. The processing chain is based on a biologically plausible model of place cells in the hippocampal system [18] but also integrates modifications that allow for better results in the case of robotic implementations [1].

The place recognition level drives the learning process in a one-shot way. That is to say, whenever none of the learned places activities is greater than a vigilance level v, new categories (signatures, contexts and places) and associations are learned. This way, the system learns independently without human supervision.

On both levels, a position in the environment is encoded as a constellation of neural activities merging "what" and "where" information – that is to say couples of visual signatures and the absolute orientations where they were observed (azimuths). The "where" information can be obtained by integrating vision and proprioception [22] or simply using a magnetic compass. *Gist* and *Landmarks* signatures encode global low-resolution and local high-resolution descriptors respectively.

The activity of each neuron g_i representing a gist signature at time t is given by the following equations:

$$g_i(t) = 1 - \frac{1}{N_{GD}} \sum_{j=1}^{N_{GD}} |w_{ij}^{GD}(t) - d_j^g(t)| \qquad (1)$$

where d_j^g is the j^{th} element of the global descriptor vector of size N_{GD} and w_{ij}^{GD} is the weight of the synaptic link between g_i and d_j^g. The learning rule of these neurons is the following:

$$\frac{dw_{ij}^{GD}}{dt} = d_j^g(t) - w_{ij}^{GD}(t) \qquad (2)$$

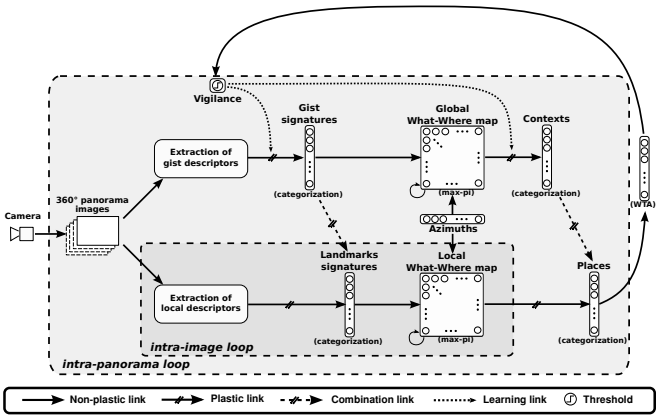

Fig. 2. Context-based model for place recognition. Two parallel neural networks extract visual features, merge them with azimuth information and categorize What-Where patterns to obtain two levels of localization respectively based on global and local descriptors. The combination link serves for the contextualization of information in the local vision pathway based on the global vision one. Learning links are only shown for the gist signatures and contexts for the sake of readability; but all the plastic links are modified based on the vigilance signal.

With this type of neurons, the idea is to save the input pattern in the links weights. Thus, the closer the input is to the learned pattern, the stronger the neurons activations. Although not biologically plausible, this method shows good generalization properties for robotic place recognition [1].

Moreover, the recognition of a landmark signature l_i not only depends on the input descriptor but also on the recognition of associated gist signatures. Thus, the activity of each neuron l_i at time t is given by the following equations:

$$l_i(t) = \frac{1}{N_{cxt}} \sum_{q=1}^{N_{cxt}} \left(\max_{k=1..N_G}{}^{q} (w_{ik}^G(t).g_k(t)) \times \left(1 - \frac{1}{N_{LD}} \sum_{j=1}^{N_{LD}} |w_{ij}^{LD}(t) - d_j^l(t)|\right)\right) \qquad (3)$$

where d_j^l is the j^{th} element of the local descriptor vector of size N_{LD}; N_G the size of the gist signatures vector; w_{ij}^{GD} and w_{ik}^G are the weight of the synaptic links between l_i and d_j^g; and g_k respectively and $\max_{i \in I}^{q}$ is an operator that gives the q^{th} maximum in I. When implemented, only the N_{cxt} most recognized gist signatures are read. The activities of landmark neurons that are not associated to them are not computed in order to reduce the computational cost.

The learning rule for w_{ij}^{LD} is the same as w_{ij}^{GD} in (2) (replacing g subscripts and superscripts by l). In contrast, the links between gist and landmarks signatures use a hebbian-like learning rule. The purpose is to capture co-activations between contextual information (global aspect of the image) and local visual input (regions of interest):

$$\frac{dw_{ik}^G}{dt} = l_i(t).g_k(t) \qquad (4)$$

If the learning process were slow rather than one-shot, the equation could benefit from a decay factor. In our initial work

based on very long simulations, learning converged to very few contextual associations [21].

In addition, the model relies on the "where" information; that is to say the orientation in which the visual feature is observed. The input is a vector in which each neuron has a preferred direction around the yaw axis. The activities of the azimuth neurons α_i are obtained after a lateral diffusion around the neuron coding for the direction of the current visual input. In our case, it corresponds to a gaussian bell with standard deviation σ_{azim}. It integrates information about the orientation of the body (robot), head (camera) and fovea (center of the visual feature).

The "what" and "where" maps (W-W maps) are second-order tensors MG and ML in which each neuron codes for a signature-azimuth couple. A short term memory stores previous activities while the visual scene exploration is still in progress (i.e. before the end a panorama). In order to reduce the computational cost, the 360^o surrounding field is discretized in N_A orientations before the computation of the tensorial product. Then, the activities of the N_{MG} and N_{ML} W-W tensors are:

$$\begin{aligned} MG(t) &= \max[(g \otimes a), MG(t-dt).(1-R(t))] \\ ML(t) &= \max[(l \otimes a), ML(t-dt).(1-R(t))] \end{aligned} \quad (5)$$

where $N_{MG} = N_G \times N_A$ and $N_{ML} = N_L \times N_A$; N_G and N_L are the size of the signatures vector; g, l and a are the signatures and azimuths vectorial representations; R a binary reset signal triggered at the end of a panorama; and $\otimes$ is the tensorial product operator.

Patterns of activities in the W-W maps code for the current location. Such patterns are categorized in context and place vectors, in which c_i and p_i neurons respectively have the following activities at time t:

$$c_i(t) = 1 - \frac{1}{\rho_{MG}.N_{MG}} \sum_{j=1}^{\rho_{MG}.N_{MG}} \max_{i=1..N_{MG}}{}^{q} |w_{ij}^{MG}(t) - m_j^g(t)| \quad (6)$$

$$p_i(t) = \frac{1}{N_{cxt}} \sum_{q=1}^{N_{cxt}} \left(H\left(\max_{k=1..N_C}{}^{q} (w_{ik}^{C}(t).c_k(t)) \right) \times \left(1 - \frac{1}{\rho_{ML}.N_{ML}} \sum_{h=1}^{\rho_{ML}.N_{ML}} \max_{i=1..N_{ML}}{}^{h} |w_{ij}^{ML}(t) - m_j^l(t)| \right) \right) \quad (7)$$

where m_j^g is the j^{th} element of the tensor MG; m_j^l is the j^{th} element of the tensor ML; w_{ij}^{MG} is the weight of the synaptic link between c_i and m_j^g; w_{ij}^{ML} is the weight between p_i and m_j^l; ρ_M is the proportion of W-W couple required for context and place recognition; $H(x)$ is the heaviside function; and $\max_{i \in I}^{q}$ is an operator that gives the q^{th} maximum in I. Like at the signatures layer, the activities of place neurons that are not associated with the N_{cxt} best recognized contexts are not computed.

Please note that the binarization of the contextual term through the heaviside function is not mandatory. The purpose is to ensure that the dynamics of place neurons activities only depends on the input pathway (instead of being reduced by the contextual factor which is < 1). This way, we do not alter the recruitment mechanism and the same vigilance threshold can be used across all the methods considered in our experiments.

Moreover, as compared to the equation (1), the factor ρ_M has been introduced for the purpose of robustness to occlusions. This parameter also compensates the absence of activation thresholds in the previous layers of the neural network. It represents an estimation of the ratio of the W-W pattern used to code for a place that is necessary and sufficient in order to recognize it (for more details, the reader can refer to a previous study [1]). The learning rules are the same as (2) for W-W inputs and (4) for contextual information. All parameters values are given in Table II and discussed in Sect. VIII.

V. Material and Methods

A. Experimental Setup

The study we present in this paper is performed on a dataset comprising three parts *A, B* and *C*. It was recorded in the area around the university of Cergy-Pontoise in France using the Robosoft ©RobuROC. The images were captured by a fisheye camera. A magnetic compass was used to acquire the orientation data. Using a magnetic compass is the easiest way to obtain this information but we could also extract it from vision, odometry and other modalities [22][23].

The dataset fits the experimental protocol used in experiments involving online learning on real robots [17]. In order to learn a new place, the robot camera captures 15 images over a 360 degrees panorama. During this process, the robot stays still in order to avoid distortions in the representation of the place. On the other hand, in the exploration phase (the rest of the time), the robot captures 7 images per panorama in order to recognize learned places as fast as possible while it navigating.

The data consist in trajectories recorded in visually different environments (see Fig. 3). *C* (23.1 meters long) simulates on-road navigation in a quite structured but highly dynamic environment (moving cars and pedestrians). *A* and *B* (20 meters long each) simulate off-road navigation. These environments are less structured and buildings are more distant. For the sake of evaluation simplicity, all trajectories are linear on x and y axis, although perturbations were induced by the rugged nature of the field. However, the camera is stabilized using a Kalman filter to limit the pitch and roll as much as possible. The robot speed is constant during exploration. The three *A, B* and *C* were concatenated for these experiments. The whole dataset includes $74 = 23 + 25 + 26$ learning panoramas (i.e. sets of 15 images captured while stopped) and $92 = 29 + 29 + 34$ exploration panoramas (i.e. sets of 7 images captured while moving). The closest distance between two possible places is $d_{learn} = 0.93 \pm 0.03$ meters in average. An exploration panorama is completed after travelling $d_{explo} = 0.71 \pm 0.01$ meters. We calculate the

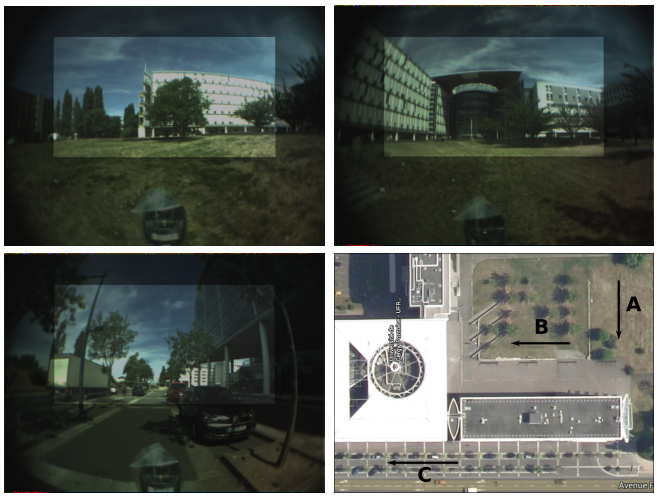

Fig. 3. Samples from the 3 dataset (A: top-left, B: top-right, and C: mid-left) and a map representing locations where they were recorded. In order to reduce distortions introduced by the fisheye lens and to remove the irrelevant bottom part, we only consider the highlighted 440×240 subimage.

ground truth by estimating the robot position at every step based on that.

All experiments are run on a 6-core 12-thread 3.33GHz CPU with 16 GB of RAM. We use the *Promethe* neural network simulator [24]. Each operation of the information processing flow (Fig. 2) can be computed as soon as the information from previous modules is updated. Independent modules are executed in parallel (i.e. in separate threads).

B. Implementation details

1) Local descriptors: We use the *rhotheta* descriptor as an input for the local vision pathway (see LPMP model description in Sect. III). First, we apply a Deriche [25] filter on the grey scale image. It consists in a derivative and a smoothing filter h and f:

$$\begin{aligned} h(x) &= c.x.e^{\alpha|x|} \\ f(x) &= b.(\alpha \mid x \mid +1).e^{-\alpha|x|} \end{aligned} \tag{8}$$

with $c = \frac{(1-e^{-\alpha})^2}{e^{-\alpha}}$. Then, the output is convolved with a DoG filter consisting in two gaussians of standard deviations σ_{DoG_1} and σ_{DoG_2}. The result is a saliency map from which we extract the points of interest.

Local views are extracted around the N_{PoI} most salient points between two disks of radius r_{small} and r_{big}. To avoid redundacies, two PoI cannot be closer that $r_{big}/2$. Then, local views are encoded using a log-polar transformation. Thereby, we obtain descriptor of size $N_{\rho\theta}$. The log-polar transformation is a biologically plausible operation, has relatively little computational cost, is invariant to small rotations and scale variations, and gives good place recognition results [1].

2) Global descriptors: As for the global vision pathway, we use subsampling-based encoding on three visual channels (luminance, chrominance and texture). Subsampling is a simple and biologically plausible process. Also, this method showed good results in the pilot study. Thus, *proflum* represents a scanline intensity profile like Milford's visual SLAM [2]. The grey scale image is subsampled at a factor of κ. Then, a 1-D vector represents the normalized sum of the pixels intensity in each column.

Likewise, *profcol* uses the same technique to associate a scanline profile to each of the chromatic dimensions of an image represented in the *Lab* color space. In this representation, *a* and *b* are color-opponent dimensions, respectively coding for the Red-Green and the Yellow-Blue axis. The *Lab* color space was designed to approximate human vision. In our case, unlike the *RGB* space, it allows for easily removing the lightness component of the image and only encodes the chrominance. Also, as compared to the *HSV* space, the two remaining color dimensions are homogeneous.

Lastly, *profgab* create a profile for each of the outputs of a Gabor filter bank. Gabor filters are defined by a sinusoidal wave multiplied by a Gaussian function and allow for orientation detection. The complex representation for a 2-D filter is the following:

$$g(x,y) = e^{-\frac{x'^2+\gamma^2 y'^2}{2\sigma_g^2}}.e^{i(2\pi\frac{x'}{\lambda}+\psi)} \tag{9}$$

with $x' = x\cos\theta + y\sin\theta$ and $y' = y\cos\theta - x\sin\theta$ where γ is the spatial aspect ratio, σ_g the standard deviation of the gaussian function, λ the sinusoidal factor wavelength, ψ the phase offset and θ the preferred orientation of the filter. In our case, the filter bank is used to detect 4 orientations (0, $\pi/4$, $\pi/2$ and $3\pi/4$).

We refer to the combination of those three global descriptors as *allprof*. The complete holistic descriptor size is thus $(7.view_W/\kappa)$ where $view_W$ is the views width. It gives a low-resolution representation of the image.

VI. Experiment 1

A. Description and protocol

In the first experiment, we consider only one pathway at a time in order to test *rhotheta*, *proflum*, *profcol*, *profgab* and

TABLE II
PARAMETERS VALUES

	Value	Description
$view_W$	440	Width of the subimages (views)
σ_{azim}	30	Std. dev. of azimuths diffusion (in degrees)
N_a	5	Nb. of orientations in the W-W Map
ρ_M	0.33	Ratio of required W-W couples (places)
N_{cxt}	7	Nb. of best recognized contexts
α	0.4	Gradient resolution (edge detection)
σ_{DoG_1}	8	Std. dev. of 1^{st} DoG gaussian (in pixels)
σ_{DoG_2}	2	Std. dev. of 2^{nd} DoG gaussian (in pixels)
N_{PoI}	5	Nb. of PoI extracted per image
r_{small}	10	Small disk radius in local views (in pixels)
r_{big}	64	Big disk radius in local views (in pixels)
$N_{\rho\theta}$	54	Size of the rhotheta descriptor
κ	4	Subsampling factor (all global desc.)
γ	0.7	Spatial aspect ratio of the gabor
ψ	0	Phase offset of the gabor filter
σ_g	16	Std. dev. of the gabor gaussian (in pixels)
λ	32	Wavelength of gabor filters (in pixels)

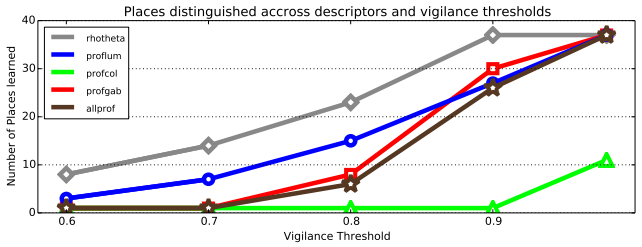

Fig. 4. Number of discriminated places by local and global descriptors for several values of the vigilance threshold. In general, global descriptors recruit less place cells than *rhotheta*

allprof separately. Our objective is to compare the descriptors capacity to discriminate places in terms of granularity. As explained in Sect. IV, a new place is learned whenever none of recognition level of the known places is higher than a vigilance threshold v. Consequently, the number of recruited place cells is a good metric for our test: The more places are learned, the finer the recognition.

We run several tests for $v \in \{0.6, 0.7, 0.8, 0.9, 0.98\}$. There are 74 learning panoramas that we feed to the visual processing chain one by one. We note that after a place is learned, an intermediate panorama has to be used to test the recognition. Thus, the system can learn 34 place at the most on our dataset.

B. Results

As shown if Fig. 4, global descriptors recruit less place cells than the local descriptor *rhotheta*. Since *allprof* is simply a concatenation *proflum*, *profcol* and *profgab* without normalization, its results not only depend of the variation observed on each individual descriptor but also on their sizes. However, for $v = 0.8$, it exhibits an averaged response.

The difference in terms of discrimination granularity can be seen for almost all of the vigilance values we considered. Here we are mainly interested in studying the learning process. Yet, it is worth pointing out that limit v values do not allow for a satisfying recognition during exploration.

VII. Experiment 2

A. Description and protocol

The purpose of this experiment is to evaluate the performance and computational cost of our model of context-based place recognition described in Sect. IV. In this model, the global vision pathway modulates the activities in the local vision pathway at two levels: gist signatures serve as contexts for landmarks and coarse localization as contexts for places. In order to assess the role of the contextualization in each of these layers, we also test two architectures where it is only done at the landmarks level or at the places level. Those three versions will be labelled **Cxt_LP**, **Cxt_L** and **Cxt_P** respectively. Cxt stands for context, L for landmarks and P for Places.

In addition, we compare the context-based models to the stacking method. Indeed, related work use non-hierarchical combinations of local and global visual features for scene recognition [12][13]. So, we test the case where all descriptors are concatenated as an input for the landmarks categorization (labelled **Stack_L**). Also, we consider the case where the gist signatures are processed separately (without modulation of the landmarks activities) but the two W-W maps are put together for place recognition (labelled **Stack_P**). In other words, the *combination links* in Fig. 2 are replaced by a simple concatenation operation and the contextualization terms in equations (1) and (6) are omitted. Lastly, the **LPMP** model (using the local vision pathway only) is used as a baseline.

Given the results obtained in the previous test, we set the vigilance threshold to $v = 0.8$.

B. Measures

In the evaluation of a visual scene recognition system, two criteria are essential: generalization and recall-precision trade-off. When measuring recall and precision, we want to make sure that the system's capacity to return relevant answers does not decrease dramatically when tuned to return as much elements as possible. It is an evaluation from the information retrieval perspective. The generalization criterion consists in the system's ability to return relevant elements in new situations by recognizing common characteristics shared with learned patterns. In navigation, such property is crucial so that the robot can correctly perform in the real world. In particular, topologically close locations should have close recognition levels. It also allows for using control mechanisms that generate smooth trajectories by averaging their responses when several situations are recognized well enough. In this experiment, we consider three measures of performance:

3WD (3 winners distances): It measures the average distance (in meters) between the 3 best recognized places and the position where they were originally learned. It characterizes the systems ability to generalize and not only recognize places at the precise location where the corresponding visual features were learned. For the sake of uniformity amongst models versions, the sum is normalized by the average distance between two learned places in each case. Thus, the results are greater or equal to 1; 1 being the best results.

MAP (mean average precision): It is a traditional compact representation of the recall-precision curves. It is the mean of the average precision at every position. Indeed, our neural network generates new place cells activities every time the W-W maps are updated by a new visual panorama. This is analogous to a new query. So we can calculate the system precision (i.e. well ranked place cells responses) depending how many place cells activity would be considered in the output. MAP scores are in $[0, 1]$ and 1 corresponds to a perfect precision.

WNR (winners-to-noise ratio): Similarly to a classic signal-to-noise ratio, it compares the level of the desired responses to the level of background noises. In this case, we consider that the most relevant information consists in the average of the 3 winners levels while the noise is the average of the remaining place recognition levels. This measure

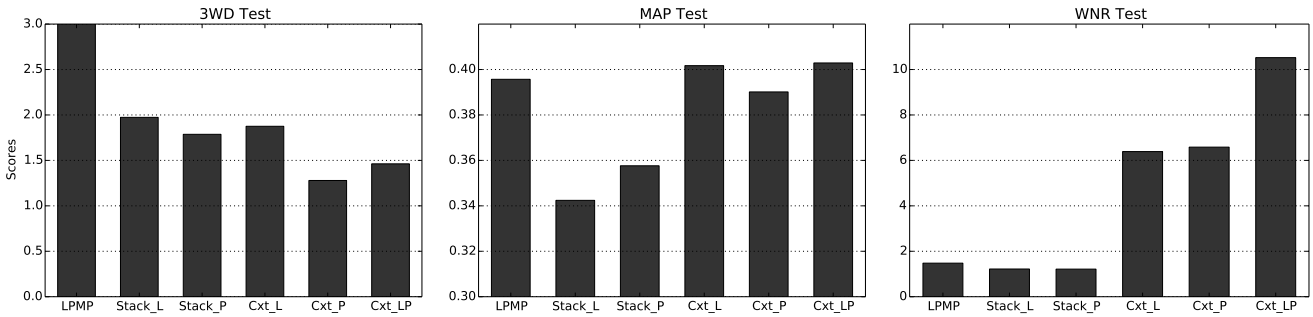

Fig. 5. Performance scores of all tested models. Our context-based model **Cxt_LP** gets the best results in almost all the tests or is second to one of its variants **Cxt_P**. 3WD: The closer to 1 the better; MAP and WNR: The great the better.

assesses whether a place cell activity decreases when the robot is far from the initial learning location. Also, higher WNR scores indicate there's less sensitivity to noise and more robustness to small variations. In other words, little risk that the fifth closest place cell would be more active wins the competition mistakenly. The WNR ratio is greater or equal to 1 and the greater the better.

Moreover, another important criterion for our evaluation is the computational cost. Indeed, more visual input to process (global+local) could make the architecture too slow to run on a real robot with real time constraints. Therefore, we measure the framerate: the number of images processed per second.

C. Results

The performance scores of the six tested models are presented in Fig. 5. First, all models combining local and global visual information outperform the LPMP model in the 3WD test. The best results are obtained by **Cxt_LP** then **Cxt_P**. Besides, context-based model get higher MAP score than stacking versions. We note that for this test, LPMP performs as well as context-based models. As for the WNR, we observe a real impact of the contextualization – **Cxt_LP** obtaining the highest score.

On the other hand, adding the global vision pathway induces an additional computational cost. Except for LPMP, the highest framerate is obtained with **Stack_P**. However, beside the absolute value, we note that the framerate stays almost constant in the case of **Cxt_LP** and **Cxt_L** while it decreases over time for all the other models.

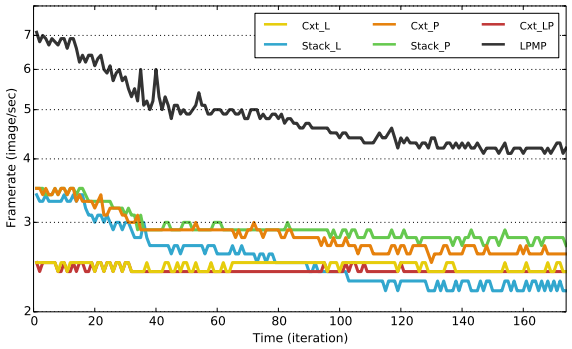

Fig. 6. Computational costs of all tested models, reflected by the framerate. For the sake of readability, the y axis is represented in a logarithmic scale. Beside the absolute value, we note that the framerates decrease over time for most of the models except **Cxt_LP** and **Cxt_L** which maintain a rather constant computational cost.

VIII. General Discussion

Most of the parameters listed in II are based on previous work and a history of experiments using the LPMP model (with the *rhotheta* descriptor) in indoor and outdoor environments [17][1]. In addition, those used to implement the global descriptor *allprof* showed good results in the pilot study.

The results obtained in Experiment 1 confirm that a global features-based system is less sensitive to visual variations than a local features-based one. Indeed, more place cells are learned using *rhotheta* than all the other descriptors. Holistic descriptors tend to capture more global information; hence the coarse granularity. Also, in the pilot study, global descriptors were able to discriminate between the A, B and C parts of the dataset (see Fig. 1). This suggests that they can be used for a coarse localization in which a set of small places can be distinguished. The context-based model we propose – consisting in **Cxt_LP** from which we derive the **Cxt_L** and **Cxt_P** for the sake of evaluation – exploits this idea.

Stack_L and **Stack_P** were implemented in order to compare the context-based model to non-hierarchical methods like in related work [12][13]. Interestingly, as opposed to our results, Lisin and colleagues obtained better performances with the stacking method than with the hierarchical one [12]. However, their experiments were based on images of multicellular organisms (planktons) on a uniform background. In contrast, we use images recorded in the outdoors in dynamic environments. This highlights the fact that robotic navigation is a specific application domain that need specific solutions.

The fact that all models using local and global visual input outperform the LPMP model in the 3WD test confirms the interests of using global descriptors for place recognition in a robotic system. Indeed, in this application domain, we want topologically close locations to have close recognition levels. However, the stacking models get lower MAP scores than the others which means that the place cells following the

3 winners often correspond to a wrong recognition. Thus, from the information retrieval point of view, these methods are less efficient than the hierarchical ones. Additionally, as expected, the context-based method proposed in this paper considerably increases the WNR by penalizing landmarks that are recognized independently from the global scene to which they are associated as well as place cells that are active outside of their contexts.

Also expected was the computational cost induced by the additional processing pathway dedicated to the global vision. Although the global descriptors are relatively compact and their computation not costly, their processing across the categorization layers of the neural network reduces the system framerate. Nevertheless, the results demonstrate this is not a critical drawback. A framerate of 2.5 images/sec means it takes 3 sec to capture a 7-image exploration panorama and update the robot localization. Since we generally set the linear speed lower than 1 m/sec in field experiments, we consider it is an acceptable framerate for a moving camera dedicated to place recognition. Additional faster sensors can be used for more critical functions like obstacle avoidance. But more importantly, beside the absolute value, we note that the framerates decrease over time for most of the models. Indeed, given that there are $N_{PoI} \times 15$ landmarks recruited per place cell, the larger the environment (or dataset) the more information has to be processed at the end of the experiment. However, the framerate stays almost constant in the case of **Cxt_LP** and **Cxt_L**. This is due to an interesting feature of our model: in the local vision pathway, only the activities of neurons associated to a limited set of well recognized ones (N_{cxt}) from the global vision pathway are computed. This is not observed with **Cxt_P** because there are not enough places learned in this experiment so that the sole contextualization of places exhibits this effect. Hence, we except that on longer trajectories, the benefit of the context-based model persists while the framerate of other techniques keeps dropping up. This should be tested in future work, as well as the impact of different values of the N_{cxt} parameter.

IX. Conclusion

Our results show that global visual features can be used for a coarse localization. In addition, in terms of place recognition, the context-based model we propose in this paper gets the best results – or is second to one of its variants. The computational cost induced by the integration of more visual information does not prevent it from being executed on real robots. More importantly, a key feature of our model makes the computational cost constant over time while it increases when other methods are used. Future work should confirm the interest of this solution on larger datasets and real robots experiments.

References

[1] C. Giovannangeli, P. Gaussier, and J. Banquet, "Robustness of visual place cells in dynamic indoor and outdoor environment," *International Journal of Advanced Robotic Systems*, vol. 3, no. 2, pp. 115–124, 2006.

[2] M. J. Milford and G. F. Wyeth, "Mapping a suburb with a single camera using a biologically inspired slam system," *Robotics, IEEE Transactions on*, vol. 24, no. 5, pp. 1038–1053, 2008.

[3] C. Siagian and L. Itti, "Biologically inspired mobile robot vision localization," *Robotics, IEEE Transactions on*, vol. 25, no. 4, pp. 861–873, 2009.

[4] T. Madl, K. Chen, D. Montaldi, and R. Trappl, "Computational cognitive models of spatial memory in navigation space: A review," *Neural Networks*, vol. 65, pp. 18–43, 2015.

[5] A. Oliva and A. Torralba, "Building the gist of a scene: The role of global image features in recognition," *Progress in brain research*, vol. 155, pp. 23–36, 2006.

[6] M. Milford, W. Scheirer, E. Vig, A. Glover, O. Baumann, J. Mattingley, and D. Cox, "Condition-invariant, top-down visual place recognition," in *ICRA 2014*. IEEE, 2014, pp. 5571–5577.

[7] K. Rebai, O. Azouaoui, and N. Achour, "Hs combined histogram for visual memory building and scene recognition in outdoor environments," 2014.

[8] P. Gaussier and S. Zrehen, "PerAc: A neural architecture to control artificial animals," *Robotics and Autonomous Systems*, vol. 16, no. 2-4, pp. 291–320, 1995.

[9] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International journal of computer vision*, vol. 60, no. 2, pp. 91–110, 2004.

[10] H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool, "Speeded-up robust features (surf)," *Computer vision and image understanding*, vol. 110, no. 3, pp. 346–359, 2008.

[11] K. E. Van de Sande, T. Gevers, and C. G. Snoek, "A comparison of color features for visual concept classification," in *Proceedings of the 2008 international conference on Content-based image and video retrieval*. ACM, 2008, pp. 141–150.

[12] D. Lisin, M. Mattar, M. B. Blaschko, E. G. Learned-Miller, M. C. Benfield *et al.*, "Combining local and global image features for object class recognition," in *Computer Vision and Pattern Recognition-Workshops, 2005. CVPR Workshops. IEEE Computer Society Conference on*. IEEE, 2005, pp. 47–47.

[13] V. Rostami, A. R. Ramli, and O. Sojodishijani, "Integration of global and local salient features for scene modeling in mobile robot applications," *Journal of Intelligent & Robotic Systems*, vol. 75, no. 3-4, pp. 443–456, 2014.

[14] M. Bar, "Visual objects in context," *Nature Reviews Neuroscience*, vol. 5, no. 8, pp. 617–629, 2004.

[15] M. M. Chun, "Contextual cueing of visual attention," *Trends in cognitive sciences*, vol. 4, no. 5, pp. 170–178, 2000.

[16] D. Navon, "Forest before trees: The precedence of global features in visual perception," *Cognitive psychology*, vol. 9, no. 3, pp. 353–383, 1977.

[17] A. Jauffret, N. Cuperlier, P. Tarroux, and P. Gaussier, "From self-assessment to frustration, a small step toward autonomy in robotic navigation." *Frontiers in neurorobotics*, vol. 7, no. 16, 2013.

[18] P. Gaussier, A. Revel, J. P. Banquet, and V. Babeau, "From view cells and place cells to cognitive map learning: processing stages of the hippocampal system." *Biological cybernetics*, vol. 86, no. 1, pp. 15–28, 2002.

[19] D. M. Smith, "The hippocampus, context processing and episodic memory," *Handbook of behavioral neuroscience*, vol. 18, pp. 465–630, 2008.

[20] M. I. Anderson and K. J. Jeffery, "Heterogeneous modulation of place cell firing by changes in context," *The Journal of neuroscience*, vol. 23, no. 26, pp. 8827–8835, 2003.

[21] N. Cuperlier, P. Gaussier, and M. Quoy, "Interest of spatial context for a place cell based navigation model," in *From Animals to Animats 10*. Springer, 2008, pp. 169–178.

[22] C. Giovannangeli and P. Gaussier, "Orientation system in robots: Merging allothetic and idiothetic estimations," in *13th International Conference on Advanced Robotics*, 2007, pp. 349–354.

[23] P. Delarboulas, P. Gaussier, M. Quoy, and R. Caussy, "Robustness study of a multimodal compass inspired form hd-cells and dynamic neural fields," in *From Animals to Animats 13*. Springer, 2014.

[24] M. Lagarde, P. Andry, and P. Gaussier, "Distributed real time neural networks in interactive complex systems," in *CSTST*, 2008, pp. 95–100.

[25] R. Deriche, "Using canny's criteria to derive a recursively implemented optimal edge detector," *International Journal of Computer Vision*, vol. 1, no. 2, pp. 167–187–187, Jun. 1987.