Visual features extraction from variable scale images using early retina mechanisms

Thibaud Debaecker, Ryad Benosman

Institut des Systèmes Intelligents et de Robotique
4 place Jussieu, 75005 Paris
{debaecker, benosman}@isir.fr

Abstract. Most of robotic systems need visual outfits to interact with the environments. Since the arrival of humanoid robots with foveolar vision, and other mobile platforms with catadioptric sensors, variable scale imaging has became a major issue to work on biologically realistic data. In this aim, the use of visual neurosciences has been limited to early vision mechanisms of cortex areas and higher order features where the camera is seen as a simple acquisition tool. It seems that at the early stages of acquisition, the retina processes a dual coding action on images. This paper proposes a new bio-inspired method to extract information from foveolar images giving a model to the lateral inhibition mechanism of the retina. The presented model shows that using a redundant pyramidal architecture produces a coding effect of the image that reduces noise. In addition it provides an implicit coding of images: the reduced original image, the location of edges and eventually texture. Experimental results on real images are presented along with a scene recognition task showing the reliability of these features.

1 Introduction

Machine vision has always used neuroscience as an inspiration for several important tasks (stereo [2], derivative filtering [3], gabor filtering [5]). Several methods have been developed to extract image features, gaussian derivative receptive fields give a local description of features [6] whereas low derivative measure the basic geometry of features [7]. It is interesting to point out that most considered features are biologically provided by high visual areas (visual areas located in the cerebral cortex called V1 and V2). Few attention has been paid to the early retina mechanism that is the first processing operated by the visual system. This paper uses the topology of the retina as a starting point to inquire on the possible processing derived from its structure. It is thought that the internal structure of the eye apart from all the known foveolar advantages [8] probably pre-processes the visual information introducing a pre-coding of the acquired images. This eases the tasks performed later by the visual cortex and other brain areas. It is also thought that there might be different channels of coding as contrast is particularly important at borders, whereas intensity is important away from them. Edges being among the most important features for segmentation of scenes, lateral inhibition might be at the beginning of image segmentation. The information provided by this step is probably non accurate but enough to be used as an entry for higher-levels areas [9]. The presented work is proposing a method performing both features extraction and noise

Workshop Proceedings of SIMPAR 2008
Intl. Conf. on SIMULATION, MODELING and PROGRAMMING for AUTONOMOUS ROBOTS
Venice(Italy) 2008 November,3-4
ISBN 978-88-95872-01-8
pp. 377-388

minimisation while preserving the information content. It relies on the use of the lateral inhibition mechanism and the foveolar structure of the retina showing their dual importance. It will be shown that the model provides an implicit image coding, giving the reduced original image, the location of edges and eventually texture.

In section 2, foveolar vision is presented followed by the architecture of the preprocessing based on a local decomposition of histograms. The mathematical model is introduced in section 3, with considerations about noise, information content, and segmentation. Finally, conclusion and future works will be found in section 4.

2 Foveolar image : spatial geometry and preprocessing

2.1 The retina simplified structure

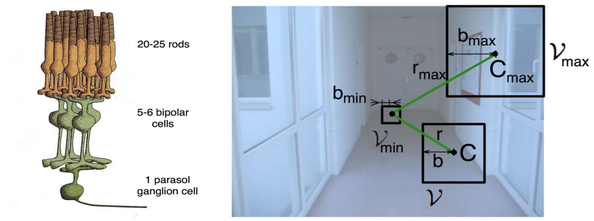
For readers not familiar with these biological mechanisms please refer to [1]. The retina is a multi-layered structure involved in signal transduction. This structure is constituted by two kinds of cells: the interneurons (horizontal and amacrine cells), which aid in signal processing, and the informative neurons (bipolar and ganglionic cells), which contain and transmit the information. Each layer contain both informative neurons and interneurons. Moreover the amount of standard neurons decreases when closer to the optic nerve, (Fig. 1(a)). And finally despite the data amount reduction, the organisation of cells are not purely convergent. There is statistically one ganglionic cell for five bipolar cells and about twenty photoreceptors, but one bipolar cell is connected to more than an only ganglionic cell. The structure is then over-all convergent and locally divergent, which brings in information redundancy. Thereby it appears that the information contained in a bipolar cell considers the information of many photoreceptors, and in similar manner, the information contained in a ganglion cell considers information of many bipolar cells.

2.2 Foveolar image decomposition

Similar to mammals visual system, the developed method takes into account the foveolar vision which can be introduced by merging contiguous variable sets of pixels using variable neighbourhoods windows in a classical perspective image as shown in Fig. 1(b). Let $\mathcal{V}$ be a square subimage and b be the half-width of $\mathcal{V}$, as shown in Fig. 1(b). The bounds of b are b_{min} and b_{max} with $b_{min} \leq b \leq b_{max}$. It is essential that the values of b depend according to the distance between C (the center of $\mathcal{V}$) and the center of the whole image. This distance is actually the r component of polar coordinates.

With r_{max} the upper bound of r, the relation between b and r is then given by :

$$b = b_{min} + \frac{b_{max} - b_{min}}{\log(2)} \log\Big(\frac{r}{r_{max}} + 1\Big). \tag{1}$$



(a) The convergent aspect of retina neurones. (b) Radius influence on the size of subimages to ensure the foveolar structure of the image.

Fig. 1. Convergent aspect and non linearity of the retina.

Image coding of foveolar areas Different kinds of visual features could be extracted from the foveolar image. But due to the non linear resolution, the accuracy of the detection can severely be altered according to the location of the feature within the image. The choice made here consists of using statistical tools (local histograms) inside foveolar areas in order to start an image coding and implicitly extract features as will be shown further.

In order to reduce the amount of information of the histogram, it is decomposed as a Gaussian Mixture Model (GMM, [10]) with Expectation Maximisation algorithm (EM, [4]). The Bayesian Information Criterion (BIC, [12]) is used to find the right number of Normal Distribution (ND) to correctly characterize an histogram. This step is applied to each RGB-level of the colour image, that enables to keep colour information.

Optimisation of histogram representation Considering a subimage $\mathcal{V}$. After GMM decomposition, the distrubution of its gray-levels is defined as the histogram $H_{\mathcal{V}}(x)$:

$$H_{\mathcal{V}}(x) = \sum_{n=1}^{Nbg} m_n \mathcal{N}_{(\mu_n,\sigma_n)}(x), \tag{2}$$

$\mathcal{N}_{(\mu,\sigma)}(x)$ is the normal distribution whose standard deviation and mean are σ and μ:

$$\mathcal{N}_{(\mu,\sigma)}(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(\frac{x-\mu}{\sigma})^2}, \tag{3}$$

and m_n is the corresponding weight of the ND n, in the global histogram $H_{\mathcal{V}}(x)$ strictly composed by Nbg ND. In most of case, one normal distribution fits with one class of pixels existing in the neighbourhood $\mathcal{V}$. The most representative ND are sorted according to their weights.

2.3 Multi-layer coding

From the acquired images to the gathering layer In the following sections, $\mathcal{V}_i^j$ is a subimage, (i is its index, and j the layer to which it belongs). The acquired image, is set to be the initial layer corresponding to $j = 0$, in what follows $\mathcal{V}^j$ will implicitly represents the whole layer. As shown an Fig. 2, subimages of $\mathcal{V}^0$ are gathered according to the foveolar decomposition: gathering layer cells receive the information of variable size pixel sets, depending on the distance from the image center. A subimage $\mathcal{V}^0_{Max}$ located in the macula encodes more pixels than the one $\mathcal{V}^0_{min}$ located in the fovea.

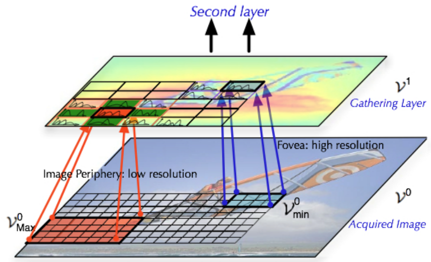


Fig. 2. Data reduction from an acquired image $\mathcal{V}^0$ to the first layer $\mathcal{V}^1$.

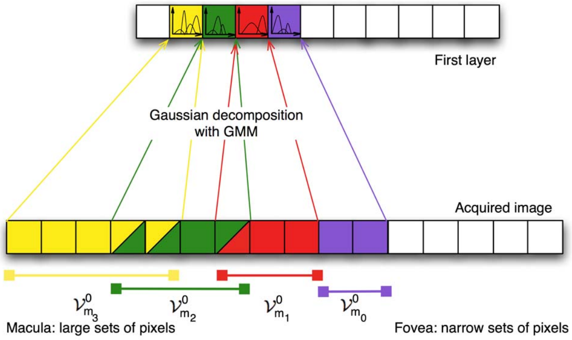


Fig. 3. Layer-1: convergent/divergent properties.

For a better understanding, assuming that the acquired images are taken from a linear camera (one rank of pixels), Fig. 3 shows a planar representation of the mechanism from the original image to the gathering layer. Four subimages are represented: $\mathcal{V}^0_{m_3}$, $\mathcal{V}^0_{m_2}$, $\mathcal{V}^0_{m_1}$ and $\mathcal{V}^0_{m_0}$. Two contiguous subimages have common pixels introducing redundancy into the system. Each subimage can be characterized by its histogram. This histogram is then GMM-decomposed to provide the information of the cells of layer-1.

From the gathering layer to the second layer The system proposed here introduces also a second layer, whose information are controlled by a process acting as interneurones. These interneurones consider the information of many neurones to inhibit or ease

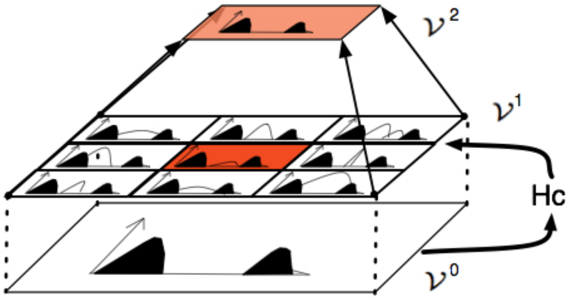


Fig. 4. Control process: the set of histograms and the control ones share the black colored ND, therefore only these two ND are transmitted to the following layer.

the transmitted data to the next layer informative neurone. As shown in Fig. 4, the GMM histogram of the subimage of $\mathcal{V}^0$ corresponding to the original image of the neighbourhood of $\mathcal{V}^1$ is used as a control information to inhibit or ease the transmitted data to the next layer. The information is stored in the GMM-decomposed H_c.

3 Multi-layer decomposition: Mathematical and Visual results

3.1 Model Generation

In the aim of obtaining a functional model, we will set in this section the mathematical framework of the decomposition operated by layer-2 in order to give a predictable model and provide a better understanding of the process. The information contained in an informative neurone of the second layer can be expressed as:

$$H_i(x) = \sum_{n=1}^{Nbg} m_{i_n} \mathcal{N}_{(\mu_{i_n},\sigma_{i_n})}(x), \qquad i = 2. \tag{4}$$

where i is the layer index, n is the index of the ND and Nbg is their number. The question is now to express the layer-2 parameters as functions of the layer-1 parameters. Modifiying the influence of a Gaussian histogram does not mean changing the averages and the standard deviations but obviously the weights of the normal distribution. The information of a layer-2 neurone is linked to a layer-1 neighbourhood. Let $\mathcal{V}_1^1$ be a neighbourhood of layer-1. We then set :

$$\mu_{2_n} = \frac{\sum_{\mathcal{V}_1^1} \mu_{1_n}}{\text{card}(\mathcal{V}_1^1)} \quad \text{and} \quad \sigma_{2_n} = \frac{\sum_{\mathcal{V}_1^1} \sigma_{2_n}}{\text{card}(\mathcal{V}_1^1)}. \tag{5}$$

Using the Bhattacharyya proximity defined from the Bhattacharyya distance [11] $\mathcal{D}_\mathcal{B}$ as:

$$\mathcal{P}_\mathcal{B}(X,Y) = 1 - \mathcal{D}_\mathcal{B} = \sum_i \sqrt{X(i).Y(i)}, \qquad 0 < \mathcal{P}_\mathcal{B} < 1, \tag{6}$$

X and Y are two same size normalized vectors, the weights m_2 are provided by:

$$m_{2_n} = \sum_{\mathcal{V}_1} \mathcal{P}_{\mathcal{B}}\Big(m_{1_n}\mathcal{N}_{(\mu_{2_n},\sigma_{2_n})}(x), H_c(x)\Big) \tag{7}$$

with $H_c(x)$ defined as equation 4 (c being reserved for control parameters). Replacing each term, it comes that:

$$m_{2_n} = \sum_{\mathcal{V}_1} \mathcal{P}_{\mathcal{B}}\Big(m_{1_n}\mathcal{N}_{(\mu_{2_n},\sigma_{2_n})}(x), m_{c_n}\mathcal{N}_{(\mu_{c_n},\sigma_{c_n})}(x)\Big). \tag{8}$$

$$\text{We set:} \quad P = \mathcal{P}_{\mathcal{B}}\Big(m_{1_n}\mathcal{N}_{(\mu_{2_n},\sigma_{2_n})}(x), m_{c_n}\mathcal{N}_{(\mu_{c_n},\sigma_{c_n})}(x)\Big). \tag{9}$$

Moreover these expressions are valid for each normal distribution of the histogram, therefore n will be voluntarily missing:

$$P = \int_{\mathbb{R}} \sqrt{\frac{m_1}{\sigma_2\sqrt{2\pi}} e^{-\left(\frac{x-\mu_2}{\sigma_2}\right)^2} \frac{m_c}{\sigma_c\sqrt{2\pi}} e^{-\left(\frac{x-\mu_c}{\sigma_c}\right)^2}}\, \mathrm{d}x \tag{10}$$

$$P = K \int_{\mathbb{R}} e^{-\frac{1}{2}\left(\frac{x-\mu_2}{\sigma_2}\right)^2 - \frac{1}{2}\left(\frac{x-\mu_c}{\sigma_c}\right)^2} \mathrm{d}x, \tag{11}$$

$$\text{with} \quad K = \sqrt{\frac{m_1 m_c}{2\pi\sigma_2\sigma_c}}. \tag{12}$$

The numerator of the exponential function is quadratic, it can then be written as:

$$P = K \int_{\mathbb{R}} e^{-\left(\alpha(x-\beta)^2+\gamma\right)} \mathrm{d}x = K' \int_{\mathbb{R}} e^{-\alpha(x-\beta)^2} \mathrm{d}x \tag{13}$$

$$\text{with} \quad \alpha = \frac{1}{2}\Big(\frac{1}{\sigma_2^2} + \frac{1}{\sigma_c^2}\Big) \quad , \quad \beta = \frac{(\mu_c\sigma_2^2 + \mu_2\sigma_c^2)}{\sigma_c^2 + \sigma_2^2} \quad , \quad K' = Ke^{-\gamma} \tag{14}$$

$$\text{and} \quad \gamma = \frac{1}{\sigma_2^2\sigma_c^2}\Big((\mu_c\sigma_2)^2 + (\mu_2\sigma_c)^2 - \frac{1}{2}\frac{(\mu_c\sigma_2^2 + \mu_2\sigma_c^2)^2}{\sigma_c^2 + \sigma_2^2}\Big). \tag{15}$$

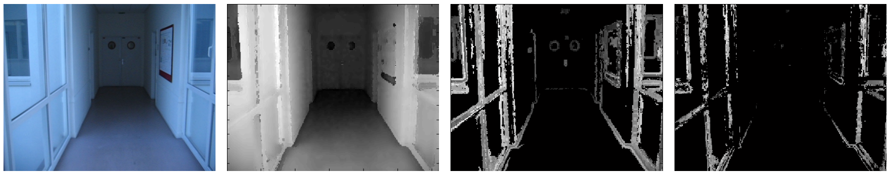
Considering the result:

$$\int_{\mathbb{R}} e^{-x^2} \mathrm{d}x = \sqrt{\pi}, \tag{16}$$

After the variable substitution $X = \sqrt{\alpha}(x - \beta)$, it comes finally that:

$$P = K'\sqrt{\frac{\pi}{\alpha}}. \tag{17}$$

This last equation means that the information of a layer-2 neurone is completely computable and predictable from the values of layer-1.



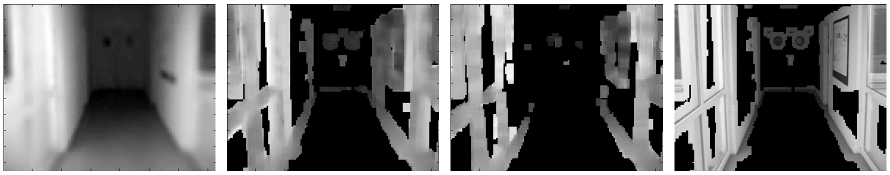
(a) Original 640×480.　(b) 1-ND.　(c) 2-ND.　(d) 3-ND.

Fig. 5. The mean images of the first four must weighted normal distribution of a GMM decomposed image in layer-1. 310 × 230.

3.2 Multi-layer image processing

First layer: Scene decomposition Fig. 5 shows the decomposition of the image appearing in Fig. 5(a) representing an indoor scene. Each processed image 5(b), 5(c), 5(d), corresponds to the ND averages of respectively the first (1-ND), the second (2-ND), the third (3-ND) ND provided by the GMM. We notice that the 1-ND image (Fig. 5(b)) is very close to the original image as most details can still be perceived even with the decrease of resolution. In the 2-ND image (Fig. 5(c)) uniform areas are set to zero. Actually this second image provides elementary edges. The 3-ND image provide more and more complex images regions corresponding to three or four (at least) pixels classes. These ND images can correspond in some cases to textured areas. It is important to notice that the whole visual features seen in Fig. 5 are still complex and non accurate. The edges defined by a wide variety of gray-levels are surely difficult to use at this step.

Second layer: Scene segmentation After the layer-2 reduction data, the information contained by the informative cells are shown in Fig. 6 where the control influence can be observed. Comparing all images of Fig. 5 and Fig. 6, the images of Fig. 6(b), and Fig. 6(c) look indeed like images of Fig. 5, but the gray level of edge portions are more homogeneous. Moreover, the edges seem to be simpler, with filled gaps. The data reduction combined to the foveolar decomposition fills the textured areas. This means that using image of Fig. 6(b) as a mask for the original image provides an automatic segmentation of the interesting scene areas (shown in Fig. 6(d)).

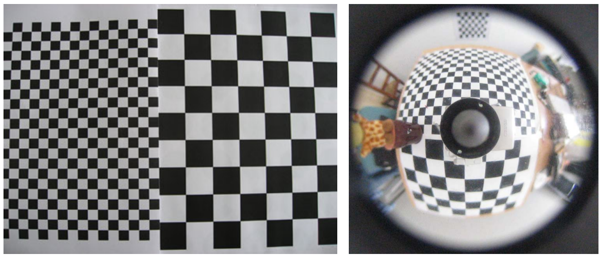


(a) 1-ND.　(b) 2-ND.　(c) 3-ND.　(d) Automatic segmentation

Fig. 6. The mean images of the first four ND of second layer-2. 145 × 105, and the automatic segmentation of discriminant scene areas.

3.3 Retinal coding and catadioptric images

Because of their non linearity, there is an analogy between biological retinas and catadioptric sensors. In both case, a picture area close to the image periphery covers a field of view wider than one located on the center.

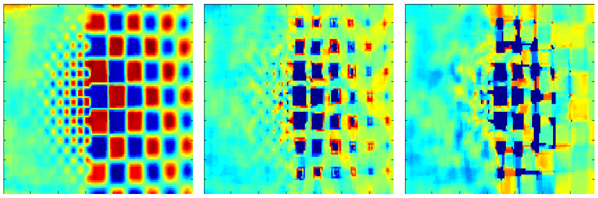


(a) Classic image. (b) Catadioptric image.

Fig. 7. Test pattern images token with classic and catadioptric sensors.

Two images (FIG. 7) are used to show the parallel between the presented coding and omni-directionnal vision. The first one is a classic image of two different size test pattern provided by a classic sensor (FIG. 7(a)). This image will be coded as described above. The other image is the same scene viewed by an omni-directionnal sensor located in the middle of the pattern (FIG. 7(b)). This image will be also coded as described above, but to take into account that this is already a non linear image, all gathering windows will have the same size, wherever the location can be. Using the notation presented in FIG. 3:

$$\mathcal{V}^0_{min} = \mathcal{V}^0_k = \mathcal{V}^0_{max}. \tag{18}$$



(a) 1-ND. (b) 2-ND. (c) 3-ND.

Fig. 8. Classic sensor coding results: 5 pixels $< \mathcal{V}^0_k <$ 30 pixels

Coding images 1-ND, 2-ND, and 3-ND of FIG. 7(a) are shown in FIG. 8.

The visual aspects of the images observed in FIG. 9 and in FIG. 8 are very close. Details in the periphery are merged together, and details located at the center are still well conserved. At the center the coding merges areas small enough to define the big square insides with an

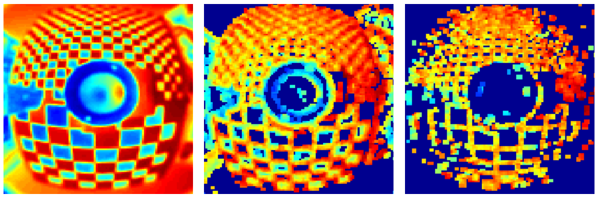

(a) 1-ND. (b) 2-ND. (c) 3-ND.

Fig. 9. Results for a catadioptric image with constant gathering windows.

only pixel class, which means only one ND. It shows that despite that there is still different geometric properties (particularly about straight line conservation), the coding proposed here extract characteristic areas which are specially information-rich, whatever the case: catadioptric or foveolar. The images provided are very close to salliency maps [14], and correspond indeed to a potential way of scan for the eye on this kind of images.

3.4 Noise independence

As introduced in section 1, one of the roles of the retina neural coding seems to be the noise reduction [13]. A polar striped image of increasing gray levels (Fig. 10(a)) has been used to study the influence of noise. Three kinds of noise are used: Gaussian additive of 20 (AG), Salt and Pepper of 10 (SP), and multiplicative of 30% (MU). Layer decomposition are applied on each image.

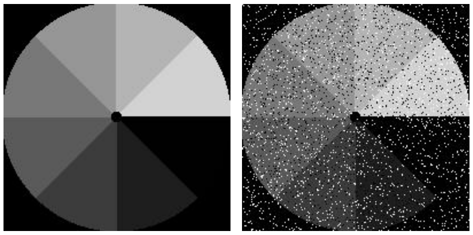
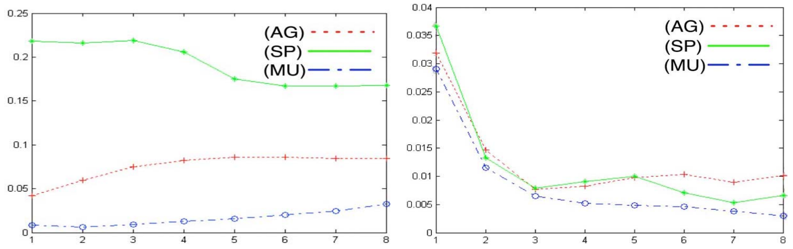

(a) Original image (b) Example of a noisy image (impulsive noise of ten).

Fig. 10. Original and noisy image before coding.

For each stripe corresponding to a particular gray-level it is possible to compute the standard deviation of the noise, using the difference between the original and the noisy image. This process provides graphics given by Fig. 11. Fig. 11(a) shows that before processing, all the standard deviations are very different. SP noise decreases with the gray-level, AG noise is almost constant, and MU noise increases. Moreover, each standard deviation mean value differs from the others.



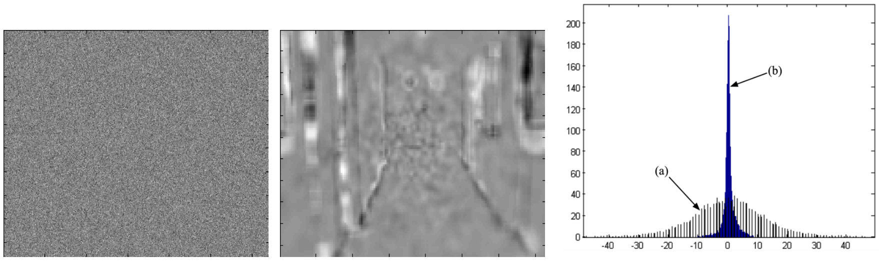
(a) Maximal ordinate value: 0,22.

(b) Maximal ordinate value: 0,037.

Fig. 11. Standard deviation of original noised and processed images according to gray values before coding (a), and after (b).

Fig. 11(b) shows that all these differences disappear after processing the image : standard deviation mean values are almost the same, and the different curves get nearly the same shape and generally decrease strongly (see ordinate axis units of Fig. 11(a) and Fig. 11(b), its mean values decrease for all types of noise: 83% for AG, 94% for SP, and 47% for MU).

Noise reduction has been tested on the image of Fig. 5(a) with an additional gaussian noise whose standard deviation is 10. Fig. 12(a) and Fig. 12(b) express the residual noises. Fig. 12(c) represents the histograms superposition of these both noisy images, before and after coding, in (a) the original image, in (b) after coding. The standard deviation of the noise decreased from 9.3 to 1.5 (84%).



(a) Residual noise before processing.

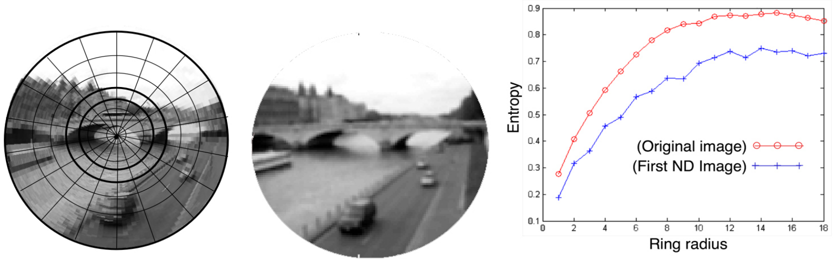
(b) Residual noise after processing.

(c) Histogram superposition of residual noises.

Fig. 12. Residual noise: Images and their histogram comparison.

3.5 Codification results: Information content

It is interesting to study the impact of the method on the amount of information contained in both original foveolar image (Fig. 13(a)) and the processed one (Fig. 13(b)). In order to extract



(a) Foveolar mapping: 480 × 480.

(b) 1-ND: 117 × 117.

(c) Full images entropy.

Fig. 13. Original and codified images and the entropy of radius-variable size ring images.

the influence of the foveolar gathering, an entropy measure is computed on 18 annular rings R_k, $k \in [1, 18]$ following a log-polar coverage as shown in Fig. 13(a). This quantity of information is given for a ring R_k by :

$$E(R_k) = -\sum_{c=0}^{c=255} Occ(R_k = c) \log P(R_k = c). \qquad (19)$$

with $Occ(R_k = c)$ the occurrence of c in R_k and P(c) is the probability of appearance of the grey value c within R_k. The entropy values for the whole image (considering I instead of R_k in equation 19) is presented in Table 1. It can be noticed that the information content value of the original image and the 1-ND image are very close (gap of 0.5%). The results using rings R_k are given in figure 13(c). A very large amount of information is conserved after processing, it is mostly located within the 1-ND images.

Table 1. Entropy values.

Original	1-ND	2-ND	3-ND	4-ND
7,12	7,08	5,57	3,32	1,01

3.6 Application: space localisation

An eye-viewpoint sequence of a man walking through corridors, gallery, stairs, along a trajectory is shown in Fig. 14. N different locations characterize this sequence (here, $N = 16$). A simple scene recognition algorithm [15] is tested on this sequence with 76 randomly chosen images. 64 location are well recognized which represents 84%.

4 Conclusion, opening, and future works

This paper presented a biologically-inspired implementation of the retina preprocessing neural coding. A functional model has been mathematically obtained enabling theoretical results confirmed by experiments. The system provides image coding, giving implicitly reliable features due

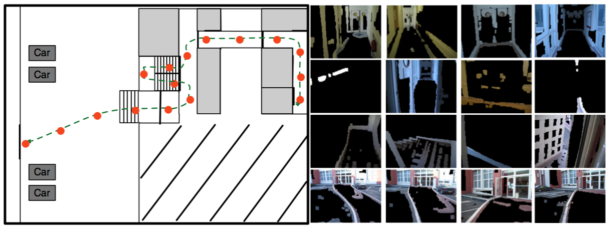


Fig. 14. Validation sequence, each image is a reference learned location.

to the robustness to different kind of noises and to the information content conservation despite the resolution decrease. A simple localisation task using these features has been tested to ensure its reliability. Providing edges and textures, and preserving the original image in the first most weighted ND, this image coding also allows a wide variety of applications.

References

1. Robert W. Rodieck, *La vision*, de Boeck Universitée, première édition, 2003.
2. Marr D. and Poggio T. A computational theory of human stereo vision. Proc. Royal Soc. London B, **204** (1979) 301–328.
3. Gabor, D. Theory of communication. J. IEE, **93**: (1946) 429–459.
4. Dempster, A.P., Laird, N.M., and Rubin, D.B. Maximum likelihood from incomplete data via the EM algorithm. J. R. Stat. Soc. B. **39**:1–38, 1977.
5. Jones J. and Palmer L. An evaluation of the two-dimensional gabor filter model of simple receptive fields in cat striate cortex. J. Neurophysiology, **58** (1987) 1233–1258.
6. Schiele B. and Crowley J. Recognition without correspondence using multidimensional receptive field histograms. International Journal of Computer Vision **36(1)** (2000) 31–50.
7. Koenderink J. and van Doorn A. Generic neighborhood operators. IEEE Transactions on Pattern Analysis and Machine Intelligence, **14(6)** (1992) 597–605.
8. Tistarelli M. and Sandini G. On the advantages of polar and log-polar mapping for direct estimation of time-to-impact from optical flow. Transactions on PAMI, **15(4)** (1993).
9. Adelson E. H. Perceptual organization and the judgment of brightness. Science, **262** (1993) 2042–2044.
10. Everitt B. and Hand D. Finite mixture distributions. Chapman and Hall (1981).
11. Kailath T. The divergence and bhattacharyya distance mesaures in signal selection. IEEE Trans. Commun. Technol. COM, **15** (1967) 52–60.
12. Schwartz. Estimating the dimension of a model. Annals of Statistics, **6** (1978) 461–464.
13. Srinivasan M. V., Laughlin S. B., and Dubs A. Predictive coding: A fresh view of inhibition in the retina. Proc. R. Soc. Lond, **216** (1982) 427–459.
14. Itti, L., Koch, C.: Computational modeling of visual attention. *Nature Reviews Neuroscience* 2(3), 194–203, 2001.
15. I. Ulrich and I. Nourbakhsh. Appearance Based Place Recognition for topological Localization. In *Proceedings of the IEEE International Conference on Robotics and Automation*, 1023 – 1029, 2000.