Modeling non-standard retinal in/out function using computer vision variational methods

Elaa Teftef, Maria-Jose Escobar, Aland Astudillo, Carlos Carvajal, Bruno Cessac, Adrian Palacios, Thierry Viéville, Frédéric Alexandre

Modeling non-standard retinal in/out function using computer vision variational methods

E. Teftef[1], M.-J. Escobar[2], A. Astudillo[2], C. Carvajal[1], B. Cessac[4], A.G. Palacios[3], T. Viéville[1], F. Alexandre[1].

(1) Inria Mnemosyne/Cortex `http://team.inria.fr/mnemosyne`**, France.**
(2) Universidad Técnica Federico Santa María, Electronics Engineering Department, Chile.
(3) Universidad de Valparaíso, Centro Interdisciplinario de Neurociencia de Valparaiso, Chile.
(4) Inria Neuromathcomp, France.

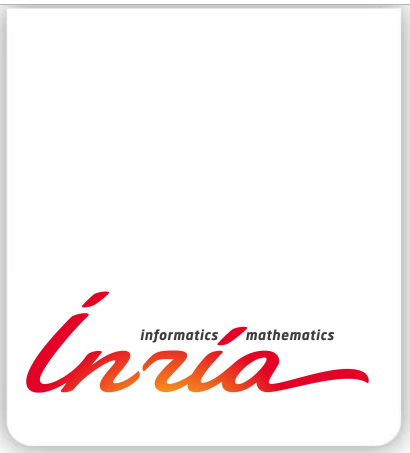
RESEARCH REPORT

N° 0000

Janvier 2013

Project-Teams Mmemosyne

ISSN 0249-6399 ISRN INRIA/RR--0000--FR+ENG



Modeling non-standard retinal in/out function using computer vision variational methods

E. Teftef[1], M.-J. Escobar[2], A. Astudillo[2], C. Carvajal[1], B. Cessac[4], A.G. Palacios[3], T. Viéville[1], F. Alexandre[1]. *

(1) Inria Mnemosyne/Cortex `http://team.inria.fr/mnemosyne`, France.
(2) Universidad Técnica Federico Santa María, Electronics Engineering Department, Chile.
(3) Universidad de Valparaíso, Centro Interdisciplinario de Neurociencia de Valparaiso, Chile.
(4) Inria Neuromathcomp, France.

Project-Teams Mmemosyne

Research Report n° 0000 — Janvier 2013 — 27 pages

Abstract: We propose a computational approach using a variational specification of the visual front-end, where ganglion cells with properties of retinal Konio cells (K-cells), are considered as a network, yielding a mesoscopic view of the retinal process. The variational framework is implemented as a simple mechanism of diffusion in a two-layered non-linear filtering mechanism with feedback, as observed in synaptic layers of the retina, while its biological plausibility, and capture functionalities as (i) stimulus adapted response; (ii) non-local noise reduction (i.e. segmentation); (iii) visual event detection, taking several visual cues into account: contrast and local texture, color or edge channels, and motion base in natural images. Those functionalities could be implemented in the biological tissues

We use computer vision methods to propose an effective link between the observed functions and their possible implementation in the retinal network base on a two-layers network with non-separable local spatio-temporal convolution as input, and recurrent connections performing non-linear diffusion before prototype based visual event detection.

The numerical robustness of the proposed model has been experimentally checked on real natural images. Finally, we discuss in base of experimental biological and computational results the generality of our description.

Key-words: No keywords

* Supported by the CONICYT/ANR KEOpS project and CORTINA associated team.

RESEARCH CENTRE
BORDEAUX – SUD-OUEST

351, Cours de la Libération
Bâtiment A 29
33405 Talence Cedex

Modélisation de la fonction d'entrée/sortie non-standard de la rétine à partir de méthodes variationnelles de vision par ordinateur

Résumé : Nous proposons ici une approche fonctionnelle de la description des propriétés des cellules rétiniennes dites Konio, ceci au niveau du réseau, en utilisant une spécification variationnelle, ce qui donne une vue mésoscopique du processus de calcul de la rétine.
Le cadre variationnel est implémenté comme un simple mécanisme de diffusion non-linéaire, mécanisme de filtrage avec rétroaction, suivi d'une couche d'unités ajustées à un élément statistique de la scène, comme on l'observe dans les couches synaptiques de la rétine pour ces cellules.
On se propose de capturer les fonctionnalités suivantes: (i) adaptation de la réponse aux statistiques des stimuli naturels, (ii) réduction non-locale du bruit (en lien avec la segmentation de l'image), et (iii) détection d'événements visuels, en tenant compte de plusieurs indices visuels: contraste local, texture ou couleur, ces amers étant généralisables à des canaux de calcul de mouvement. Ces fonctionnalités peuvent être mises en œuvre dans les tissus biologiques, comme on le discute ici.
Nous utilisons des méthodes de vision par ordinateur pour proposer une description fonctionnelle du calcul effectué au niveau de la rétine.
La robustesse numérique du modèle proposé a été vérifié expérimentalement au niveau numérique sur de véritables images naturelles. Nous discutons, sur la base de résultats expérimentaux biologiques et informatiques, la généralité de notre description.

Mots-clés : Pas de motclef

1 Introduction

Recently it has been proposed that the retina, a accessible part of the brain, sustain more complex behaviors than expected before (Masland & Martin, 2007), including spatial, temporal and motion recognition (Schwartz & Michael, 2008) (Olveczky, Baccus, & Meister, 2003). The retina correspond to a multi-stream device including from phototransduction to early visual processing and neural coding, at different spatial and temporal scales. At the application level, a computing parallel information streams system for (e.g.) visual prosthesis is likely to need an important post-processing level before visual signals feed the nervous system (Barriga-Rivera & Suaning, 2011). For the later, an adequate computational architecture should have at least two-layers network with non-separable local spatio-temporal convolution as input, and recurrent connections performing non-linear diffusion before prototype based visual event detection.

The present article is organized as follow: In the section 2, we review key biological facts at the early dorsal and ventral K-cells visual streams, computing sophisticated spatial and temporal pattern recognition as review in the next section, 3, base on (Gollisch & Meister, 2010; Litke et al., 2004; Schwartz & Michael, 2008; Olveczky et al., 2003; Sterling, 2004). In the subsequent section 4, we implement the computational principles raised by the study on the retinal K-cells (Hendry & Reid, 2000; Yoonessi & Yoonessi, 2011) using a variational specification of the visual front-end, base on a network of retinal ganglion cells. In the section 5, we implemented a numerical model to study real image sequences and finally model predictions are presented to verified biological validity.

2 From standard to non-standard early-vision front-end

Standard front-end retinal streams

The retinal output correspond to several different spatio-temporal processing, base on a diversity of ganglion cells (Gc) types, the output of the retina to the nervous system, of the incoming image sequence. Such streams are embodied as separate strata that span the retina at the Gc surface.

It is generaly accepted, that the (i) parvo (P-cells) and (ii) magno (M-cells) streams, compute respectively (i) color image details and contrast in central vision and (ii) monochrome intensity temporal variation, including at very low contrast, in peripheral vision and (iii) konio streams (K-cells) projecting to the LGN (middle layer) receive information from blue cones, sending information to the V1 color blobs (Hendry & Reid, 2000) and from small bistratified Gc having a surround which is sensitive to yellow (G. Field et al., 2007). Here, we are going to focus on a subset of K-cells, beyond their color property, as detailed in the sequel. For other types of Gc see, e.g., (Callaway, 1998).

At the modeling level, information streams (i) and (ii) are well represented by LN models, i.e., a one layer spatio-temporal filtering followed by a static non-linearity, as represented in the right part of Fig. 2, this unit being tuned by a gain control mechanism, as reviewed e.g., in (Wohrer, 2008). Qualitatively, such non-linear filtering tends to remove spatial correlations in the visual world, producing a less redundant output, as required to transmit information in the optic nerve fibers of limited spatial capacity (see (Pitkow & Meister, 2012) for a discussion), resulting in a sparse coding for retinal spike trains in response to the statistics of natural images (Simoncelli, 2003).

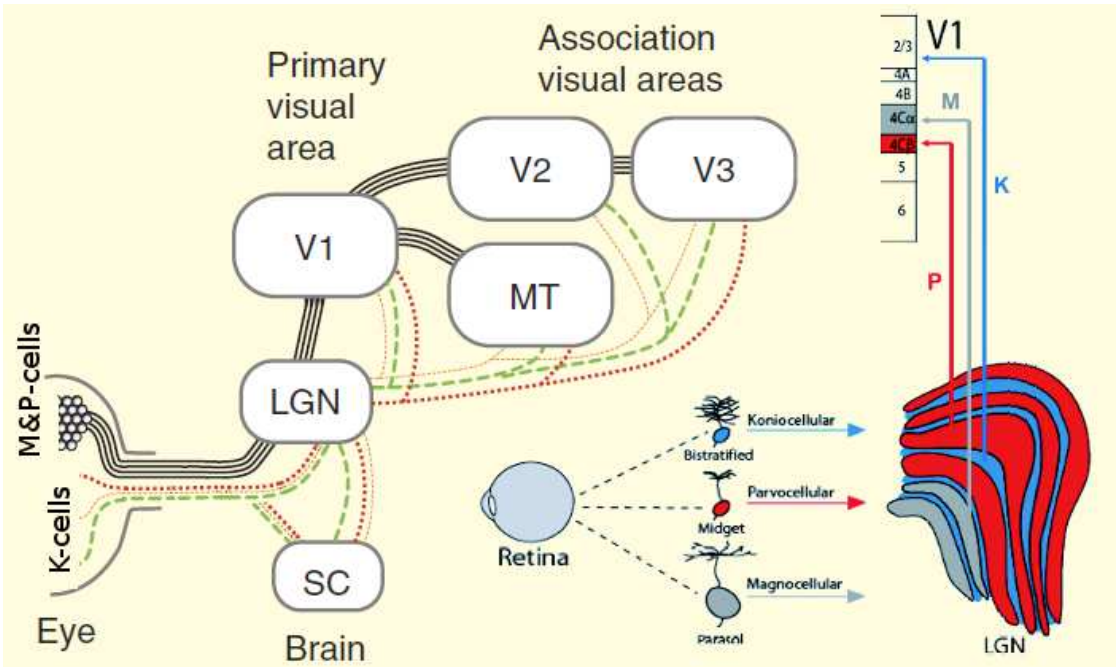


Figure 1: Schematic representation of projections of the K-cells from (Masland & Martin, 2007) and (Nassi & Callaway, 2009).

Considering retinal output K-streams

While standard magno/parvo Gc downstream correspond to the standard visual dorsal/ventral streams (leftward drawing) after a relay in the ventral/dorsal LNG layers (rightward drawing); K-cells projections are more miscellaneous (leftward drawing) and interact with the parvo/magno streams in the LGN and via the V1 2/3 (rightward drawing). Connections from the retina are feed-forward, while all other brain connections include feed-backs. In a purposive view of visual perception, the parietal cortex performs from such input an unconscious, efficient, specialized and rapid processing of the whole visual field, and prepares actions adapted to the characteristics of the environment. In addition K-cells connect via the LGN to important multi-sensorial areas like the amygdala lateral nucleus (ALN) involved in the control of emotion (Ciocchi et al., 2010), thus in direct link with survival actions.

Importance of the K-streams

Quantitatively, 80% of the Gc are midget towards the P-cells in the thalamus, because of the need of fine spatial resolution to perceive visual details. Only 10% are parasol Gc towards the M-cells in the thalamus, and 10% are bistratified Gc (K-cells) towards the superior colliculus and the LGN layers (G. D. Field et al., 2007). Though such non-standard retinal cells are in proportion few in primates (whereas up to 75% in phylogenetically earlier mammalians) they constitute robust and well-identified cell mapping of the visual field (Gauthier et al., 2009). This lower proportion is easily explained: Having large receptive field (RF) about 10 *deg*, the cover of the whole retinal field required less units.

In the human eye, about 130 millions of photoreceptors, where 120 millions are rods and 10

millions are cones, concentrate onto about 1.5 millions of Gc, including the 10^5 K-cells considered here.

The key computational aspect of K-streams

The current computational assumption here consider that K-cells streams provide rough but fast event or object -detection of visual events (Hendry & Reid, 2000)(Masland & Martin, 2007) in order to induce survival goal-directed actions on time (V. Lamme & Roelfsema, 2000), and to drive higher-level iterative processes with prior information (Callaway, 1998; Koivisto, Railo, Revonsuo, Vanni, & Salminen-Vaparanta, 2011). Given natural image sequences, we interpret these two functionalities as image segmentation for visual objects detection and natural statistical recognition, including temporal pattern recognition. In the other hand, K-cells participate to blind-sight (V. A. Lamme, 2001; Cowey, 2010), i.e., or not consciously seen, including, slowly moving targets detection, rough localization of stimulus. This is typically information not for "seeing" but for visually guided behavior. There are some evidences, that such computation starts in the retina:

(i) Some retinal Gc are able to produce sophisticated spatial and temporal pattern recognition on their own (Gollisch & Meister, 2010; Werblin, 2011), including motion detection, directional selectivity, local edge detection, object motion and looming detection.

(ii) During phylogenetical evolution, the eye was not a simple feed-forward system: Feed-backs from the remainder of the nervous system had been able to produce adaptive learning of sophisticated visual functions "optimizing the transfer of information" (Sterling, 2004).

(iii) Fast stimulus categorization develops roughly 150 ms after stimulus onset (Thorpe, Fize, & Marlot, 1996). In that respect the visual processing need to perform this using only few computation steps (Viéville & Crahay, 2004).

(iv) From an information processing point of view, regarding downstream computation, it is optimal to take into-account the whole spatial and temporal resolution of the photo-receptors before the necessary compression occurring in the optical nerve to implement such complex non-linear filtering (Simoncelli, 2003).

These are the key elements that differentiate the K-cells from parvo/magno streams. In nutschell, the K-stream provides fast, a-priori event detection assumptions, to the remainder of the visual system (V. Lamme & Roelfsema, 2000).

3 Retinal architecture and biological constraints

What are the biological constraints regarding what can be computed in the retina (i.e., the computational characteristics)?.

In order to derive a mesoscopic model of K-cells input/output function let us collect known facts about the computational properties of the retina, as schematized in Fig. 2.

Here are the three major features we propose to take into account:

1. *Not one but two layers*: It is clear that, by no means, a simple feed-forward filtering layer, even non-linear, can compute complex visual cues such as those reviewed above. On the other hand, it is known that a two layers non-linear feed-forward network, i.e., with a "hidden" layer, is a universal approximator of any function. For a static input/output relationship, i.e., a unique image as input with a unique related output, a neural network

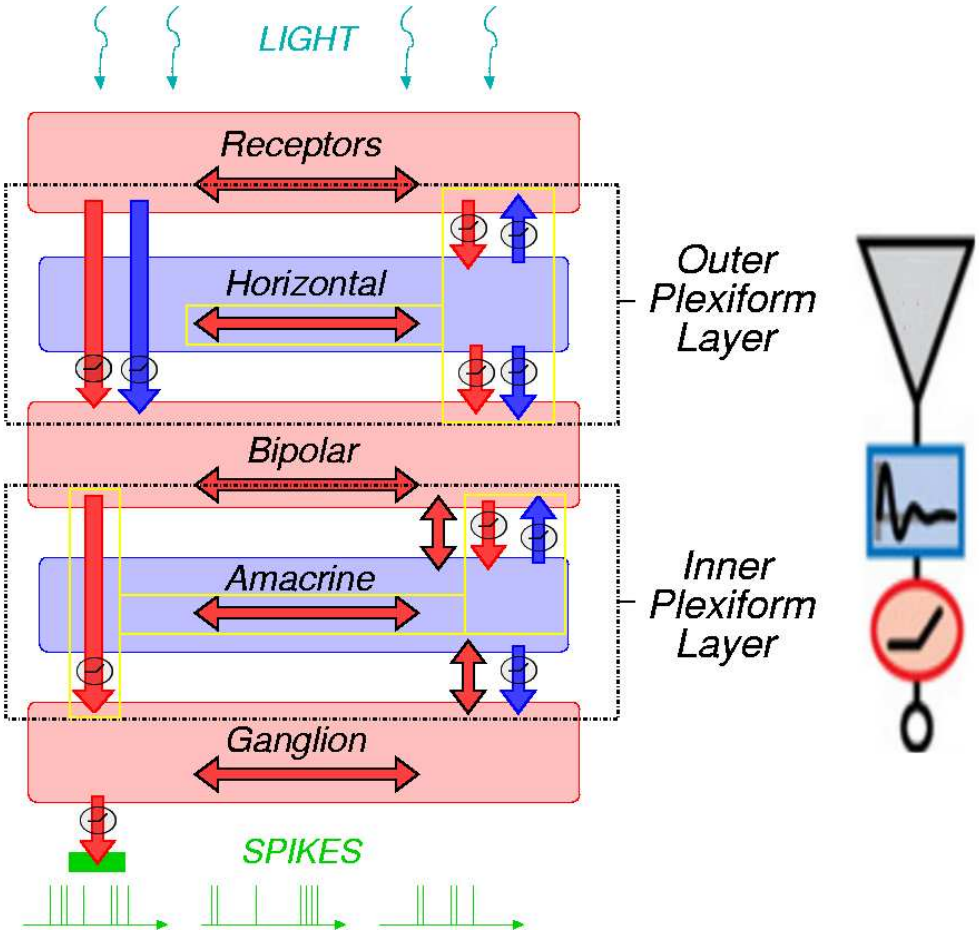


Figure 2: *left*: Schematic representation of the retinal architecture, after (Wohrer, 2008) and (Gollisch & Meister, 2010). Excitatory connections are in red, inhibitory in blue. Gap junctions are in red with a black contour, neighboring cells of the same type being generally linked through gap junctions. The direct pathway from light receptors to Gc is a two-layers process defining two successive non-linear filtering stages with different excitatory-inhibitory patterns. The present mesoscopic model is compatible with such an architecture, but *requires* a subset of this architecture (highlighted in yellow). This pathway is modulated by the interstitial cells in each layer. Here, we have emphasized that non-linearities are present on each link. *right*: The basic view of cell unit processing: (i) spatial filtering (gray arrow) specified by the wired connectivity, combined by a (ii) band-pass temporal filter, thus including delays, (blue rectangle) and (iii) followed by a thresholding operation.

(e.g. retina) with two layers is a sufficient architecture for generic computations (Hornik, Stinchcombe, & White, 1989). For a dynamic input, i.e., an image sequence, this basic idea has to be generalized, and we propose the following.

2. *Non separable spatio-temporal filtering*: At a given time, we consider that the retina input state is a function of the short-term visual information and that the output is a causal function of this 2D+T image volume.

   Quantitatively, we may consider the last 150 ms interval as temporal depth (Li, 1992), with a visual minimal event-time of about 10 ms: This thus correspond to a dozen of "frames" when time is discretized.

   - This means that the spatio-temporal filtering is local, i.e., only spatial connections from neighborhood cells are taken into account, while in the temporal domain the filtering is only due to delays in the biological substrate and connections. This seems to be fairly true for the outerplexiform layer (OPL) (i.e. for horizontal cells) which corresponds to the filtering mechanism considered here. This is less obvious for amacrine cells in the IPL (Werblin, 2011), likely involved in mechanisms not addressed here.

   - This means that the spatio-temporal filters are constrained: not all "weight" values can be used, but only values with signs compatible with the excitatory/inhibitory connections, and value ranges taking into account the nature of the connection (e.g., gap junction versus synaptic junctions). All synapses have a delay, increasing as a function of the connection length.
   - This means that the output at a given time is a static function a 2D+T volume, thus computing local motion in the image sequence, but no long-term motion cue. Therefore the previous argument of a two-layers architecture is sufficient for complex function approximation still stands for dynamic stimuli in this restrained framework.
   - This also means that the spatio-temporal filtering is non-separable: In other words, we do not consider a sequence of 2D image (thus spatial filtering each image and then temporal filtering the result) but the filtering of the 2D+T volume.

   In fact, as made explicit before (Gollisch & Meister, 2010; Dong, 2001), the retinal structure induces a separable spatio-temporal filtering if we consider a cell "alone" in a standard parvo-stream, but different if we consider interactions between cells, and also consider that the retina is adapted to natural image stimuli, as suggested by (Dong, 2001).

3. *Recurrent connections*: The general architecture definitely shows that the hard-wiring provides feed-forward connections and recurrent (horizontal and feed-back) connections, as made explicit in Fig. 2.
   - *Recurrent connections as local diffusion.* These connections are local, continuous and with tiny delays (i.e. based on gap-junctions on local non-spiking very short dendrite/axon connections) (Sterling, 2004). Qualitatively, this corresponds to local diffusion of the state value, here mainly corresponding to the membrane voltage value.
   - *Static non-linearities everywhere.* Another key aspect is the fact that each connection is subject to a static non-linearity with a main characteristic: Threshold corresponding to a rectification of the signal. The rectification is not related to action potential spiking (except for the Gc output), but to bio-chemical membrane mechanisms.

3.1 Functional characteristics of the retinal network

The retina is able to compute local isotropic contrast and intensity temporal variation detection, but requires specific functionalities.

-A- stimulus adapted response

From raw light intensity event detection (including dim light flashes), to RF responses corresponding to the basis element of natural images decomposition (Hyvärinen, 2009), the primary aspect of the retinal input-output function is to be adapted not only to the general statistics of natural images, as observed in the LGN projections (Dan, Atick, & Reid, 1996), but also to dynamic changes in the statistics of a given environment (Hosoya, Baccus, & Meister, 2005). Such stimulus adaptation is implemented by a suitable choice of the local 2D+T convolution kernel induced by the local network connectivity (Wohrer, 2008). Such filter elements span the space of possible relevant local input (Hyvärinen, 2009).

However, beyond the hard-wiring of such connections, a local adaptation must occur to optimize the response. This also means that the retinal filtering is not only adapted to "any" natural image sequence, but likely also adapted to a "the" present visual environment, as observed at the LGN level (Truccolo, 2001).

A way to interpret this, is to generalize the contrast adaptation mechanism proposed in, e.g. (Wohrer, 2008), to second-order local statistics adaptation. This means to assume that the retinal processing is able to tune not only the "gain" but also the local spatio-temporal correlations to a given environment and to react (i.e., to detect) a specific second-order spatio-temporal statistics in the signal.

This is the first main computational assumption of our work, with two consequences:

1. On one hand, as illustrated in Fig. 3, second-order statistics is in one to one correspondence with the magnitude of the (here: 2D+T) spectrum[1]. Detecting second-order spatio-temporal statistics means detecting a given spectrum magnitude signature. On the reverse, the spectrum phase is a function of the higher-order and second-order statistics, so that assuming second-order processing means that the retinal adaptation and detection capability is *only* related to the spectrum magnitude signature. This is due to the Wiener-Khintchine theorem that relates the signal auto-correlation, thus the second-order moments, the power spectral density. It has been shown that, in fact, this is a relevant cue to characterize natural image categories (Torralba & Oliva, 2003), e.g., discriminate between human-made environment or not, categorize a given landscape, etc. We would like to hypothesize that what is true for "large scale" natural image categories is also true for "small scale" image categories, e.g., detect stone vs vegetation, etc. This is plausible since the natural image statistics power spectra is $1/f$, meaning that the second order statistics is scale invariant, which suggests (but it is not established) that the same cue may be relevant at a local scale. This is going to be verified numerically in section 5. This is not obvious and this does not means that the whole statistics is scale invariant. It is true, in average, for natural images (Simoncelli & Olshausen, 2001). It is not the case for a given subset of images and there are evidence at least in V1 that even if the statistic is preserved our visual system treats images differently (Simoncini, Perrinet, Montagnini, Mamassian, & Masson, 2012).

2. On the other hand, tuning the second-order but not the higher-order statistics has a very precise meaning with respect to stimulus decomposition. It means that the retinal processing is related rather to a "Principal Component Analysis" (PCA) which decomposes

[1] see the Wiener-Khintchine theorem for a formal equivalence

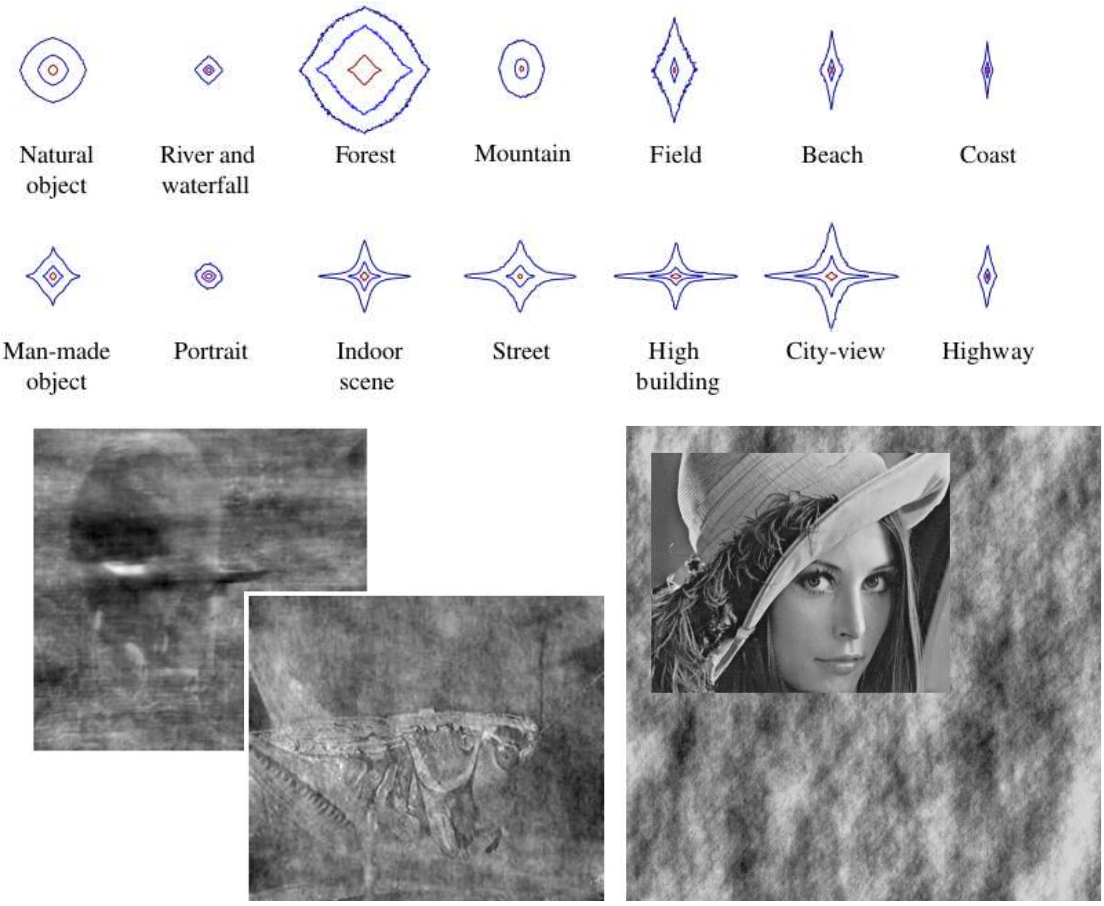


Figure 3: *Top view* : Relevant representation of natural image categories from the spectrum magnitude signature (here the 60, 80 (in blue) and 90% (in red) level curves), from (Torralba & Oliva, 2003): Each environment class has a relatively specific signature. In a bio-plausible framework such signature correlation can be computed by a linear/non-linear network subset, since the distance to a given spectral signature is equivalent to the energy response of a filter with spectral property corresponds the given signature, as made explicit in the sequel. *Bottom view*: The "semantic" of the spatial spectrum magnitude versus the spectrum phase: When (on the left, from (Hyvärinen, 2009)) corrupting the magnitude while preserving the phase, the image (here of a bull and an insect) is still recognizable, whereas (on the right, for the well-known "Lena image") when canceling the phase and preserving the magnitude the image has still "natural image statistics" but the shape has vanished. Here, we propose as a minimal model, to consider that K-cells processing is based on second-order information, thus on spectrum amplitude, since in one-to-one correspondence[1].

a signal with respect to its correlations, thus second-order statistics, than an "independent component analysis" (ICA) which decomposes the signal with respect to higher-order moments.

The fact that early vision optimized for either only a second-order statistics or for higher-order statistical dependencies is an open question and recent results tends to show that V1 RF function not only make their outputs as independent as possible, but also facilitate the processing of higher-order statistical regularities (Karklin & Lewicki, 2006). Similarly, it is assumed that high-order statistics are taken into account in redundancy reduction (Barlow, 2001).

However, results suggest that adaptation in the retina system is not sensitive to changes in commonly measured higher-order statistics, retinal cells exhibit remarkably invariant behavior to changes in higher-order statistics (Tkačik, Ghosh, Schneidman, & Segev, 2012).

This invariance is coherent with what has been observed in V1 with pink-noise, i.e., the fact it is sufficient to "imitate" second-order statistics of natural images to obtain responses similar to natural image statistics (Kayser, Salazar, & Konig, 2003). To this debate, we bring the following argument: If natural images categories are essentially characterized by their spectrum magnitude (Torralba & Oliva, 2003), thus second-order momenta, it is sufficient to detect visual events related to such second-order cues.

No mistake, this statement only concerns the K-cell processing considered here, whereas analyzing natural images using ICA at the retina level is another important issue considered elsewhere (Keskes, 2011). And we are going to numerically verify that it is sufficient to use such cues to perform the complex early-vision processing that is attributed to K-cells. We also wonder, if, since the eye is expected to produce visual input from natural images anyway, it is relevant not to transmit the obvious information that the image is natural, but the information remainder. How could it be, in that case, that though invariant with respect to higher-order momenta (such as kurtosis) the retina still "computes" higher-order momenta? In our framework, this is due to another processing step : Non-local filtering, now described and which is a non-linear transformation of the image, so that second-order and higher-order momenta are mixed, thus higher-order momenta indirectly taken into account, even if the basic detection mechanism is only due to second-order cues.

-B- non-local noise reduction

Since the main biological constraint of the retina in terms of information transmission is the 10^2 compression factor for P-cells up to 10^3 for other Gc cells in the optical nerve, reducing the noise with respect to the signal is a major issue (Simoncelli & Olshausen, 2001). The noise added by the photo-transduction mechanism is obviously to be filtered (Wohrer, 2008). However, what is to be considered as pertinent signal vs distracting information or noise in the incoming visual flow is task dependent. For instance, the background motion induced by ego-motion is a "noise" with respect to object motion recognition but a relevant signal with respect to gaze stabilization (the opto-kinetic reflex for instance has the background motion as main input cue).

Furthermore, there is a "local" notion of noise (i.e., a random spurious additive signal, locally correlated in space and time, or not), but regarding the visual perception, there are other notions: A notion of "outliers" (i.e., spurious piece of signal not to be taken into account to avoid bias, e.g., a spurious light reflection) and a notion of "inliers" (i.e., piece of signal coherent with the ground base data but inducing a mis-perception, e.g., a shadow.

The concept of noise reduction is no more related to a local filter, but to non-local operators, since the context matters. In other words, filtering means estimating a given zone in the image,

thus to perform image segmentation. This non-local idea of filtering corresponds to the grouping of homogeneous information, reducing small or spurious non-significant variations, and then produce a reduced representation in which outliers are eliminated because they're not homogeneous with respect to the information neighborhood, while inliers are grouped in another region to be further treated as signal or noise depending on the perceptual task. Such mechanism of synchronization is well-known in the retina (Gollisch & Meister, 2008; Shlens, Rieke, & Chichilnisky, 2008) where the shared noise from photoreceptors is a major cause of correlated firing (Greschner et al., n.d.), so that synchronized Gc outputs correspond to activity with the same stochastic signal characteristics.

Furthermore, it is important to point out that we do not group regions of homogeneous "intensity" but regions of homogeneous response to the 2D+T filtering discussed above, e.g., of the intensity, intensity variation, contrast, local edge or motion pattern, etc. It is important to remark that time, here, is considered only as a local feature. As a consequence, such mechanism should be able to segment on complex local spatio-temporal patterns. In particular, as explained in (Gollisch & Meister, 2010) after (Hochstein & Shapley, 1976), texture motion, even if the average local intensity remains constant, is locally detected: During texture motion, different intensity-variation cells respond at different time to different spatial patterns at their RF locations, showing the same activity profile, inducing a cumulative effect related to strong rectification (summing only positive responses in the neighborhood with both excitatory and inhibitory responses). It is a perspective of this work to numerically verify that texture motion is also compatible with such variational approach (Olveczky et al., 2003) (Aubert & Kornprobst, 2009).

To further understand this aspect of retinal processing, we are going to formalize in details to which extents the retinal architecture is able to not only smooth but also segment the visual signal in order to provide a multi-scale representation of the input in terms of regions of homogeneous signal values (upper-scale), with a coding at finer-scale of the region indexing for a given unit. In this framework we thus only consider the spiking rate and phase, and are going to code them by two analog signals, the former being a positive value, the latter being a relative (i.e., only defined by the difference with another value). Since phases synchronize, we obtain only a discrete set of distinguishable values and this is equivalent to an indexing.

-C- visual event detection

The retina is able to detect sophisticated visual events such as:

(i) Fine spatial information, as observed qualitatively with natural image sequences, a Gc being often either silent or sparsely by deterministic firing for a given image category (Soo, Schwartz, Sadeghi, & Berry, 2011);

(ii) local approaching motion, which is known as being computable by a local divergence filter (Subbarao, 1990), i.e., the weighted sum of dedicated local masks.

Such functionality includes anticipation, either motion extrapolation or omitted stimulus response (Schwartz & Michael, 2008), these complex computations coming about through the interplay of rather generic features of retinal circuitry (Gollisch & Meister, 2010). In both cases, as expected, spatio-temporal pattern yields adaptation (delay reduction in the former case, response habituation in the later) and an unexpected event induces a strong reaction. The time window order of magnitude is the second. This is thus not directly derived from a 2D+T filtering mechanism (its temporal window being too short) but by a local fading memory of previous input, thus using regressive feed-backs, and the simpler model we propose here is a simple auto-regressive mechanism. The key point is that this induces a switching mechanism, the detection of a visual event modifying the local parameters on the retina. With (Gollisch

& Meister, 2010) we assume that the switching "may be initiated by much more subtle image features", precisely by the detection of a given visual event.

Regarding the detection of such fast visual event, it has already been investigated in detail how Gc and subsequent layers can produce early responses to a visual event (Rullen & Thorpe, 2001; Thorpe & Fabre-Thorpe, 2001). We assume here, that if such a mechanism is able to produce a detection at early stages in the visual system, it is likely produced by a large contribution of the retinal processing. This is plausible because the corresponding bio-plausible most efficient computational method is a one layer method implementable a simple non-linear filter with threshold effect (Viéville & Crahay, 2004). This is related to the so-called Support Vector Machine (SVM) algorithm, and the bio-plausible implementation is a simple competition between units coding similarities with respect to prototypes. This may be implemented as a two steps process: correlation with prototyping responses and competition between the outputs, as implemented in a standard one-layer architecture with lateral inhibitory connections, which is the case of the OPL.

The statistical learning theory and the related SVM algorithm (Viéville & Crahay, 2004) allows us to assume that only a *little* set of prototypes (the support vectors) has to be recruited for the detection of an event. Moreover, "approximate" prototypes are sufficient, since the prototype of a category can be wrong as far as it is less wrong than a prototype of another category. It is thus reasonable to assume that this is a biologically plausible functional model of the retina, since it is able to detect rather complex spatio-temporal visual events. As already pointed-out, the major argument is that it is a pertinent to use the full retinal resolution to perform it before the optical nerve spatial compression.

This view is also coherent with the observation that in the presence of natural stimulus Gc have a sparse and deterministic behavior. Sparse because they react in terms of detection of not a given visual event. Deterministic in the sense that they fire a spike response at reproducible times. This is also coherent with observations already reviewed here, that the response is not limited to a "local visual field", since there is no significant visual events at such small case, but at higher sizes of the visual field. When considering the objects size in a standard natural scene (?, ?), the fact that the K-cells response corresponds to events in a $10deg$ of the visual field seems properly tuned to natural visual events (Hendry & Reid, 2000).

4 Variational specification of the visual front-end

Let us now "translate into equations" the previous elements, as developed elsewhere (Viéville & Crahay, 2004; Viéville, Chemla, & Kornprobst, 2007) and adapted to the retinal processing.

4.1 Non-local IPL filtering

The main mechanism is illustrated in Fig. 4, where non-local filtering is precisely defined from the well-known work of (Mumford & Shah, 1985), and following here the presentation of (Aubert & Kornprobst, 2009). A step further, following (Gérard, Kornprobst, & Viéville, 2007), we consider the well-defined discrete approximation of the Mumford-Shah criterion proposed by (Chambolle, 1995; Chambolle & Dal Maso, 1999). This allows us to implement this mechanism, taking the retinal IPL architecture into account, as follows. Over a 2D array of "pixels", at a resolution h,

it writes:

$$min_{\mathbf{O}} \underbrace{\sum_{ij} |\bar{\mathbf{W}} * \mathbf{I}(\omega_{ij}) - \mathbf{O}(\omega_{ij})|^2 +}_{\text{Input filtering}}$$

$$\underbrace{\sum_{ij, d\omega \in \{(0,1),(1,0)\}} \min\left(\alpha\, |\mathbf{O}(\omega_{ij}) - \mathbf{O}(\omega_{ij} + d\omega)|^2, \beta\, h\right)}_{\text{Region \& border optimization}} ,$$

with $\mathbf{I}(\omega_{ij}) = \int_{\omega \in Pixel_{ij}} \mathbf{I}(\omega)/h^2$ being the integrated value of a pixel. It has been shown by (Chambolle, 1995) that this discrete criterion is a well-defined discretization of such a function. Contrary to other numerical schemes (Aubert & Kornprobst, 2009), its distributed implementation corresponds to a simple-layer network computation with feed-backs. Another technical aspect is the fact that here we consider not a scalar but a vectorial image, and in, e.g., (Viéville et al., 2007), it has been derived how such mechanisms trivially generalize to the multi-channel case. More precisely, the minimization is implemented by the following non-linear iterative filter:

$$\mathbf{O}(\omega_{ij})_{t+1} \leftarrow \mathbf{O}(\omega_{ij})_t + \delta\, \nabla \mathbf{O}(\omega_{ij})$$

with:

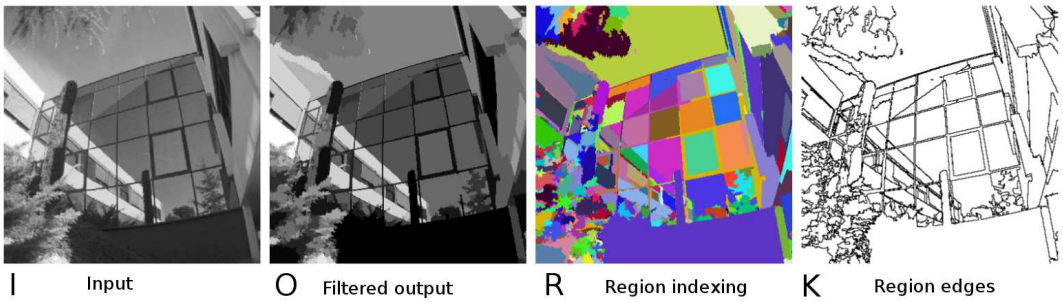
$$\begin{array}{rcl} \nabla \mathbf{O}(\omega_{ij}) & \equiv & \bar{\mathbf{W}} * \mathbf{I}(\omega_{ij}) - \mathbf{O}(\omega_{ij}) \\ & + & \sum_{d\omega \in \{(0,1),(1,0)\}} \left\{ \begin{array}{ll} 0 & \text{if } |\mathbf{O}(\omega) - \mathbf{O}(\omega + d\omega)|^2 > \frac{\beta\, h}{\alpha} \\ \alpha\, (\mathbf{O}(\omega + d\omega) - \mathbf{O}(\omega)) & \text{otherwise} \end{array} \right. \end{array}$$

.

Making explicit this recurrent equations allows us to relate it to the IPL retinal architecture discussed previously. It is thus an iterative filter, implemented as a local feed-back with *recurrent connections* and is based on *local diffusion*, as made explicit in Fig. 2 and discussed in section **??**, regarding recurrent connections. The input gain is positive ("excitatory connections") and the local diffusion (influence of the neighborhood activity) is positive too. On the reverse, the local feedback is negative, again as predicted by our knowledge of the IPL architecture. The proposed iterative filter is based on a piece-wise linear feedback, with a very precise *static non-linearity*: if the local contrast magnitude $|\nabla \mathbf{O}(\omega)|^2$ is higher than a threshold, diffusion is canceled. This is easily implemented by local inhibition, triggered by a contrast threshold, again compatible with elements of Fig. 2. This is the reason why we propose that the IPL filtering stage may be considered as implementing such non-local filtering.

4.2 Visual event detection in the OPL

As far as the choice of the *non separable* 2D+T convolution kernels are concerned, beyond usual contrast and image intensity variation kernels, we thus assume in this modeling that natural image local categories can be related to a spectral magnitude signature, i.e., normalized spectral magnitude profile. We thus hypothesize that retinal computational units are tuned to a given spatio-temporal spectral response $\bar{\mathbf{W}}_n(f)$. This means that a given input $\mathbf{I}$ corresponds to the pattern $\bar{\mathbf{W}}_n$, if its spectrum $\mathbf{I}(f)$ is closed to $\bar{\mathbf{W}}_n(f)$, up to a scale factor. Let us write $|\mathbf{I}|^2 = \int_\omega |\mathbf{I}_n(\omega)|^2$ and assume without loss of generality that the spatio-temporal spectral response is normalized, i.e. $|\bar{\mathbf{W}}| = 1$. Let us also consider the "orthogonal" pattern, i.e. the normalized profile $\bar{\mathbf{W}}_n^\perp$, with $|\bar{\mathbf{W}}_n^\perp| = 1$ and $\int_f \bar{\mathbf{W}}_n^\perp(f)\, \bar{\mathbf{W}}_n(f) = 0$, this profile being well-defined up to



$$\min_{\mathbf{O},K} \underbrace{\int_{\omega\in\Omega-K} |\bar{\mathbf{W}} * \mathbf{I}(\omega) - \mathbf{O}(\omega)|^2}_{\text{Input filtering}} + \alpha \underbrace{\int_{\omega\in\Omega-K} |\nabla \mathbf{O}(\omega)|^2}_{\text{Region homogeneity}} + \beta \underbrace{\int_{\omega\in K} 1}_{\text{Border parsimony}}$$

Figure 4: The computational principle of the Mumford-Shah image segmentation variational approach for a multi-channel image, after (Aubert & Kornprobst, 2009). Given an area of pixels Ω and an input image $\mathbf{I} : \Omega \to \mathcal{R}^3$ which associate to each pixel its color (here a red-green-blue (RGB) value), the goal is to compute an output image $\mathbf{O} : \Omega \to \mathcal{R}^C$, segmented in regions, writing K the border between regions. The specification is three-fold. (i) *Input filtering.* The output image is closed to the local filtering of the incoming image by 2D+T convolution kernels parameterized by the weights $\bar{\mathbf{W}}$ and computing C channels such as contrast, intensity variation, or local spectral property. (ii) *Region homogeneity.* Inside a region we want the output value variations to be minimal (here ∇ stand for the gradient operator, thus the value variation), yielding homogeneous regions. (iii) *Border parsimony.* Between regions, we want the border length to be as small as possible to draw smooth and minimal edges between regions. Minimizing the weighted sum of these three criterion allows us to specify an image segmentation, tuning the fine/coarse grained segmentation with β and tuning the region scale with α. The success of the method comes from the fact that, under realistic conditions, the optimization converges towards a piece-wise a constant intensity profile, i.e., perfect homogeneous regions. Implementing this minimization, from the derivation of the criterion gradient using the related Euler-Lagrange equation, leads to a distributed algorithm with precise local weights adaptation rules (Viéville et al., 2007). Region indexing (pixels of the same region have similar output values) and edge detection (edge pixels have non negligible local contrast) is also obviously provided by such process.

a negligible measure for the given dot-product. Applying usual equalities we can relate this mechanism to a simple convolution with normalization and static non-linearity as follows:

$$\bar{\mathbf{W}}_n(f) \simeq \lambda\,\mathbf{I}(f) \quad \Leftrightarrow \quad \epsilon \simeq \int_f |\frac{\bar{\mathbf{W}}_n(f)}{|\bar{\mathbf{W}}|} - \frac{\mathbf{I}(f)}{|\bar{\mathbf{I}}|}|^2 = \int_\omega |\frac{\bar{\mathbf{W}}_n(\omega))}{|\bar{\mathbf{W}}|} - \frac{\mathbf{I}(\omega)}{|\bar{\mathbf{I}}|}|^2$$
$$\Leftrightarrow \quad \left\{ \begin{array}{l} \frac{\bar{\mathbf{W}}_n(\omega) * \mathbf{I}(\omega)\big|_{\omega=0}}{|\bar{\mathbf{I}}|} \simeq \int_f \bar{\mathbf{W}}_n(f)\,\bar{\mathbf{W}}_n(f) = 1 \\ \frac{\bar{\mathbf{W}}_n^\perp(\omega) * \mathbf{I}(\omega)\big|_{\omega=0}}{|\bar{\mathbf{I}}|} = \int_f \bar{\mathbf{W}}_n^\perp(f)\,\bar{\mathbf{W}}(f) = 0 \end{array} \right.$$

This obvious textbook derivations make explicit two facts: if spectral magnitude signatures are similar, this simply means that the two normalized signals are similar, while using the reference signal as a normalized convolution (convolution after normalization of the input signal) yields a unitary response.

Local spectral signature detectors are thus easily implementable in the biological substrate as a convolution with a normalized input signal, as it is the case for a simple neural cell with gain control, a static non-linearity providing the thresholding mechanism after the convolution, to cancel negative or negligible parts of the signal. A step further, if two coupled convolution kernels are coupled with their "negation", i.e. with kernel representing the orthogonal response, the visual event detection can be based on more robust mechanisms of "margin" comparison, detecting the event if the "positive" channel response is significantly higher than the "negative" one.

5 An effective implementation of visual events detection

In order to address the question "How is it possible for the retina 'alone' to efficiently compute, on natural image sequences, such sophisticated visual responses", we have developed a software (Teftef, 2012; Teftef & Viéville, 2012) that implements the previous described functionalities. This piece of code is based on the CImg[2] image processing open-source library, and is itself a piece of open-source code available under the K-mrs[3] nick-name. For the generic segmentation routine, we reuse the Léonard Gérard implementation (Gérard et al., 2007) of the related Chambolle algorithm (Chambolle & Dal Maso, 1999), and have contributed to CImg on this aspect. The event detection SVM mechanism has been implemented using the open-source libsvm[4] of Chih-Chung Chang and Chih-Jen Lin.

As far as software engineering is concerned, the key point is that we work on vectorial data volumes, i.e. an "image" is a data structure with pixels indexed by their 2D location (x, y) and temporal depth t, each pixel being a vector $\mathbf{c}$ of channels, starting with the red, green, blue channels, as input, extended with computed channels (e.g., spectrum signature). As a consequence, all functions developed here are usable for both static and dynamic event detection, and all related processes.

While the numerical verifications proposed here only consider static images, spatio-temporal convolutions and other operators are available for the software to process 2D+T images sequence slices.

The SVM learning of the spatial event detection prototype was a complete standard process: We have selected images of variable sizes of each category, and run the calculation of the intensity channels and spectral channel, as shown and explained in Fig. 7. Given the red (R), green (G) and blue (B), original channels we have chosen the redundant coding : G-R, G-B, G in addition to the intensity R+G+B. The only reason was to obtain numbers rather stable, given the data

[2] `http://cimg.sourceforge.net`

[3] `M-mrs, for KEOpS mesoscopic retina simulator http://project.inria.fr/keops/K-mr`

[4] `http://www.csie.ntu.edu.tw/~cjlin/libsvm`

Figure 5: Schematic representation of middle-scale visual events as a support-vector-machine (SVM) implemented in terms of a competitive winner-takes-all connectivity. Here, areas of 3 to 10 deg of the visual field are considered with, as examples, "flowers" (upper-left), "gravel-floor" (middle) and "vegetation" (upper-right). The discrimination between such regions is based on the trivial competitive response of the different detectors. Such winner-takes-all mechanism is obviously implemented with spike-coding (Rullen & Thorpe, 2001) since the higher the response, the sooner the spike generation, which can inhibit other responses and thus "win" the competition. The key point, for the present process, is the choice of the "prototypes". As pointed out in (Viéville & Crahay, 2004), the very powerful statistical learning theory, leading to the SVM algorithm, can be revisited as a competitive comparison between prototypes. These prototypes, are the support vectors, close to the hyper-surface that separates two categories. The key parameter is the "margin" between these prototypes and the frontier between two categories: The higher the margin, the more robust the detection on the subsequent data set (generalization property). In the present numerical experimentation, a standard SVM supervised mechanism have been used.

set intensity distribution, with for instance, a predominance of green, while red or blue coloration of the image is often relative to green zones. We recall that the filtering is not linear, so that even if the SVM algorithm is invariant by an affine transformation of the features space, the segmentation step depends on such coding choice.

Obviously, the representation is easily adaptable, for instance to dichromate visual systems, like in rodents, using B+G or UV+G, so somehow different from trichromats vision (Peichl et al., 2005). As soon as these two colors allow to discriminate the natural scene objects, the previous framework applies. This is in coherence withj the fact that the retinal visual system is adapted to the natural environment of the animal (Sterling, 2004).

This event detection is performed on the non-local filtered images (in fact it has been qualitatively experimented that event detection does not work on the raw images in this context).

The choice of considering the sum of the spectrum amplitude responses at 50% of the maximal spectrum response is a pure experimental choice, as detailed in Fig. 7, but guided by two considerations. On one hand, this can be implemented by a simple linear / non-linear filter, as follows: each point of the spatial spectrum corresponds to the correlation of the signal by the corresponding trigonometric functions. The maximal spectrum response always corresponds to the continuous (or DC) component of the spectrum, i.e., the average intensity value. Tuning, in a given 2D direction, the spatial frequency in order the response to correspond to 50% of the DC component, is a simple gain control feedback. Finally, computing the signature reduces to a simple summation. By no means we assume here that this is what happens in the retina, nor that this is the only solution. In fact, as developed in the previous section, the biological substrate offers a much more generic mechanism, computing the correlation between the observed and expected spectrum amplitude.

Though, as discussed previously, all functions can be implemented as distributed computations on a biologically plausible neural network, for obvious computation time performances issues, we have used optimized code in this numerical implementation. Previous studies on these aspects (Viéville & Crahay, 2004; Viéville et al., 2007) have extensively addressed the issue of both the segmentation and detection implementation in a biological network.

The originality of the numerical experimentation is the fact that we have considered a rather specific set of visual stimuli, as made explicit in Fig. 6, and involving studying a specific animal model (see (Delgado, Vielma, Kähne, Palacios, & Schmachtenberg, 2009; Ardiles et al., 2012) for more details on these aspects) [5]. Restraining the issue to computer vision process only, it means that we have considered "degraded visual conditions" (with respect to state of the art image sequence data base), the issue being to check whether and to which extents such processes "still work" in such restrained conditions.

In order to experiment in a rather tricky context, we have applied the proposed framework to the ability to recognize shapes in the image (Masland & Martin, 2007), i.e. restraining to 2D+T volumes with a thickness of one. We thus have implemented the fact that we consider that the retina is able to compute sophisticated second-order visual cues, including spectrum amplitude signature, since this is implementable as a linear/non-linear filter, see Fig. 7. We have implemented the non-local filtering of the different visual streams as a mechanism of image segmentation, see Fig. 8 for results. Finally, considering that the retina is able to detect static or time-varying visual events, we propose after (Viéville & Crahay, 2004) a suitable mechanism for such biological computation, evaluated on a realistic data set, see Fig 9. More technical details are available in (Teftef, 2012). In particuler, this report includes comparisons with different types of filtering mechanisms in order to illustrate the benefits of the non-local filtering proposed in

[5] As an example, an interesting issue regarding the will to work with such "rodent specific visual data", is to wonder whether the rodent visual system is "optimized" for such visual environment, or if on the contrary such rodent visual system is phylogenetically conserve with respect to such ecological variation.

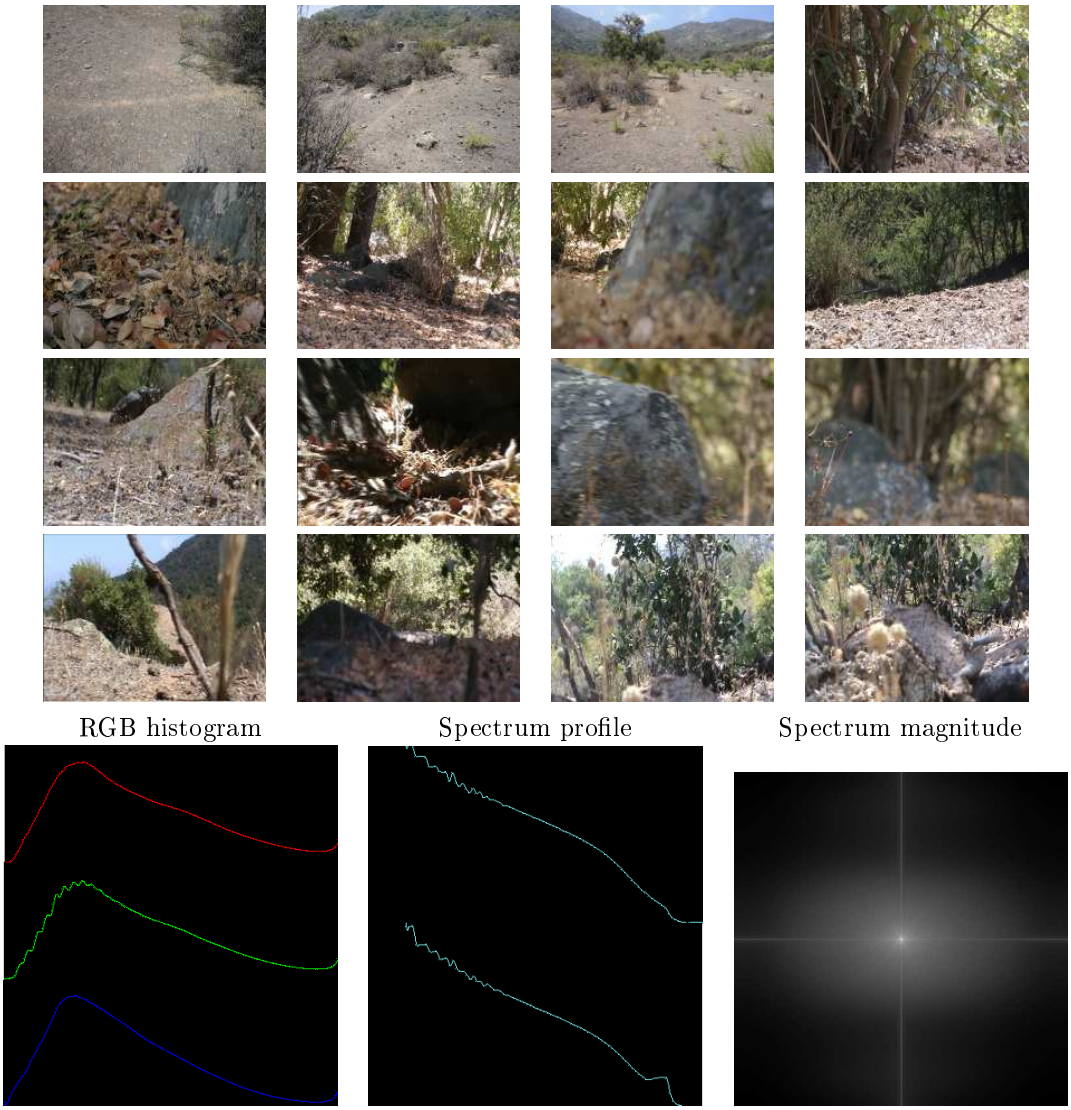


Figure 6: The "La Campana database: a natural habitat of Octodon degus rodent" (40 sequences of 10^2 images of about 800×800 pixels). In order to be able to perform both biological and numerical experiments on degus visual system, image sequences of about 10^2 images each have been taken, in the natural environment of the animal, at the height, lighting conditions, image resolution and with displacements roughly corresponding to the rodent usual displacements. It is important to note that this species use to live in a specific visual environment: In the bottom-left view, the red-green-glue (RGB) intensity histogram is not Gaussian as it is usually the case (Simoncelli & Olshausen, 2001), but biased, especially in the green channel (oscillation in the distribution), due to the statistics of the vegetation colors. In the bottom-central view, two spectrum magnitude profiles in log coordinates, taken over hundred of frames, show that the $1/f^d, d \simeq 1$ expected profile (thus a line with a negative slope in the figure) does not occur here, whereas the spectrum is a function of such visual context.

Figure 7: Considering a multi-stream computational framework, including spectrum amplitude signature. For five different zones of natural images labeled on the left picture, the spectrum amplitude and the 50% level curves are drawn, from the non-local filtered images. From left to right : 1. a flower, 2. a piece of sky with an artifact, 3. a tree zone, 4. a dark zone, and 5. a third tree zone perturbed by a piece of sky. Interesting enough is the fact that zones 3. and 5., despite their very different visual aspect have a rather similar spectrum signature, different from 1. Though zones 2. and 4. have spurious spectral responses (for 2. this is the artifact in the sky that yields the spectrum and for 4. the spectrum is not relevant in such a dark zone), since the system is also using the local color as a cue, this yields no mistake. We numerically observed that not only the whole image second-order statistics (Torralba & Oliva, 2003), but also very local second-order statistics seem to be relevant for natural image categorization. More precisely, in the present numerical implementation limited to static images, the color and the sum of the spectrum amplitude responses at 50% of the maximal amplitude are taken as cues to characterize the local texture, in order to perform region detection. The fact that such very simple scalar parameter calculated on the spectrum magnitude leads to relevant results (see Fig. 9) is a good argument in favor of the proposed choice. The generalization to spatio-temporal volumes, thus 2D+T spectra, is straightforward and is a direct perspective of this work.

Figure 8: Qualitative examples of non-local filtering using diffusion mechanisms as discussed in (Viéville et al., 2007) regarding their biological implementation. Lower images are the non-local filtering output of one channel (here intensity) computed on the corresponding upper image. Interesting enough, several key points can be observed here : The output is morally an "artificial image" corresponding to the "natural" input, where the forms in the image have been preserved. More precisely, the edges (e.g., the flowers boundaries) are preserved even when the image is "smoothed" : The fact the mechanism filters the noise and not the edges, is simply due to the non-linearity of the diffusion operator, since small variations diffuse and are thus filtered, whereas higher intensity variations related to edges do not. Furthermore, diffusion is anisotropic, more precisely performed along the edge, but not across it (in a more mathematical language : Tangential to the local edge orientation, but not orthogonal to it), thus preserving it. In the present computer implementation, region synchronization is implemented by the propagation of a "phase index" as expected in a biological neural network. In the sky, the artifact (likely a tree limb) has disappeared, because one non-local effect of this process is small enough so such regions are absorbed though the non-linear diffusion. In a nutshell, we have here a "full resolution", but simplified image.

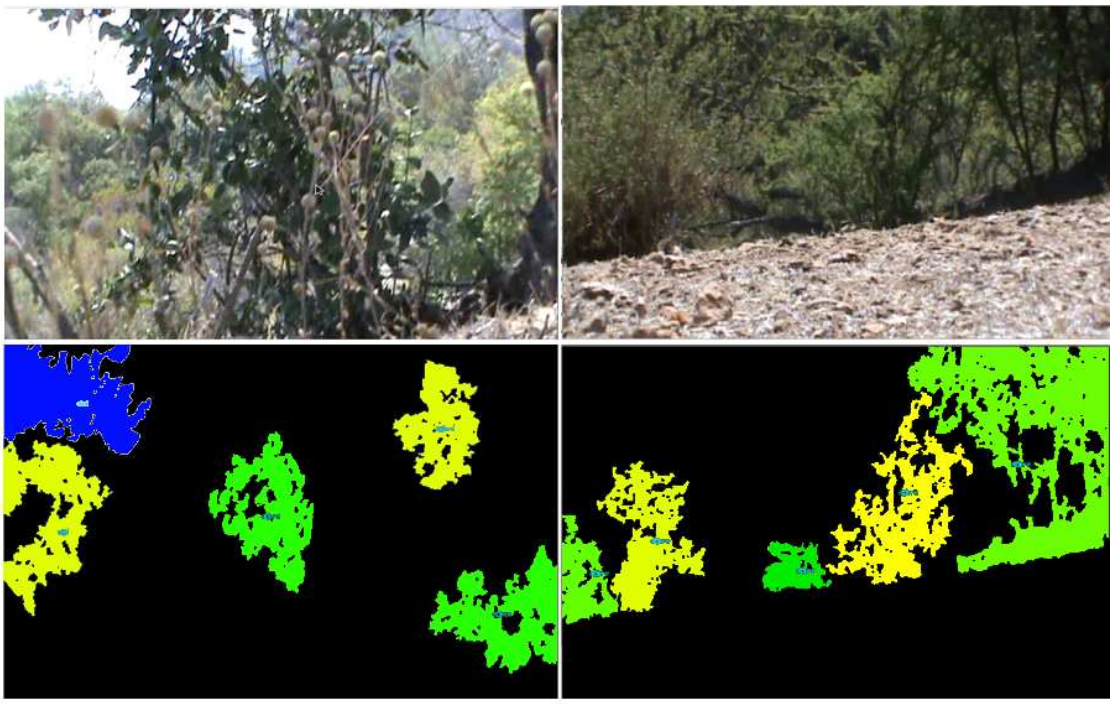


Figure 9: Two examples of detection of visual events (here, spatial events only). Top views correspond to the input image, bottom views to the detection output. The most salience regions are shown in both cases, i.e. those with maximal size and/or channel average values (here, for the present figures, size only is considered). *Left view* A "sky" zone and four dark/light "tree" zone have been detected. *Right view* The main "tree" zone have been detected, while the ground is less salient since not structured in terms of "regions". Clearly this kind of information is relevant to directly detect zones with potential pray/predators "hidden in the green". The categorization algorithm has been trained with a very small learning set (a few dozen of samples) and the recognition test is no more than a comparison of the region synchronized parameters (color, spectral signature) with respect to prototypes in competition. Such functionality is easy to generalize to another visual events (e.g. snake crawling, predator approaching motion, etc..) and thus provides an effective early-vision generation of a-priory information. We have numerically verified that such detection is robust along the image sequence. However, we also experimented that the non-linear filtering (i.e., segmentation) step is mandatory: Without this sophisticated early-vision step, the present categorization has very poor performances. This output represents sparse responses of the retina in the presence of specific visual events.

this article.

6 Discussion: What is predicted by such a modeling ?

With respect to usual models regarding the retina processing, the present work simply proposes to instantiate usual ideas about the so-called non-standard Gc behaviors, but with a computational framework that make them compatible with realistic image processing. Furthermore, this framework raises some new predictions about retinal processing that are not yet tested, but could be worth tested to better understand the biological behavior of the retina.

Here are experimental some facts that will falsify the present framework, regarding the K-cells behavior considered here.

About adaptation

We consider K-cells stream and speculate that the related processing mainly involve second-order statistics, because this is sufficient to characterize "natural image categories". It may thus be a second-order issue (Simoncelli & Olshausen, 2001) (Torralba & Oliva, 2003). A simple way to falsify the present statement would be to study to which kind of stimulus statistics a K-cell in the LGN is sensitive to. We expect such cell to be invariant to global pink noise whitening (e.g., to response to an artificial image with the same features a natural one except its spatial spectra magnitude).

About segmentation

We claim that the IPL filtering not only "reduces local noise" but performs an implicit clustering of visual zones, likely based on non-linear diffusion mechanisms yielding synchronization between cell responses of the same class. As a consequence, the compression process is not only based on very local spectral or statistical properties, but tuned to non-local cues. This means that the context must strongly influence the local response. More precisely, a Gc response must be averaged and synchronized with respect to Gc with similar responses (i.e., cells of the same "region") but not with cells of another one. If we can observe, for instance, un-synchronized cells corresponding clearly to the same value of the corresponding visual cue in the same image area, the present model is lapsed.

A key falsifiable consequence of the present assumption is the fact that the Gc channels output "image" (i.e., 2D map) must *not* correspond anymore to the statistics of natural images with a scale invariance of a spectrum in $1/f$. This is simply incompatible with the non-local region segmentation mechanism. If such statistics are preserved, the proposal is defeated. In other words, the decoding of the LGN transmitted image to the brain must have the look of an "artificial image" more than a natural one.

A less falsifiable consequence on this topic is the fact that high spatial frequencies are filtered, but not in a homogeneous way: inside a region high-frequency are expected to be canceled, information diffusion leading to a uniform responses of cells averaging the response value, whereas on the border between regions, no filtering occur. This is a weaker argument because, anyway, as the spatial frequencies are higher on edges than between edges. Here the inhomogeneity in terms of spatial high-frequency filtering must have been amplified by the IPL.

About detection

We finally have considered the visual event detection to be based on a "prototype competition". This means that a given visual event must be represented by a few spatio-temporal patterns and if the input is close to one of these patterns, the Gc detection must fire. If, on the contrary, the representation of a given visual event (e.g. a crawling motion) is fully distributed, as when represented by additive information or dynamic attraction, the proposal is to be deprecated. Another key aspect is that such prototype must represent event and no-event cases, since the underlying algorithm compares two alternatives to trigger an event. In other words, the mechanism is a comparison with respect to a ground state, not a threshold. Furthermore, we expect border-line prototypes rather than "typical" prototypes, to be coded for the detection. This means that we expect the local spatio-temporal activity to be correlated with responses that, for different categories, are close together. This is again due to the fact that the system does not have to detect an event in absolute way but relatively to an alternative. Therefore, it is more important to be able to characterize the data with respect to prototypes close to the border of such alternatives (notion of "support prototypes"), than to confirm that the similarity with a "caricaturistic" prototype. If such characteristic is detected at the experimental level, the proposed framework becomes plausible, otherwise it remains questionable.

At the mesoscopic level, i.e. considering the response of a given retinal stream at the network level, it implies that a single cell type may serve quite different roles (e.g., approaching motion and structure motion), because it is recruited for a different event detection mechanism, depending on the current state of the network. This has already been partially observed, but if this observation is marginal, whereas the majority of cells are dedicated to a unique function, the model fails.

A less falsifiable consequence is that we have to assume that the observed switching between local scale functions must be initiated by much more subtle image features, i.e., the visual event context. If such switching is coherent with the related event detection (i.e., switch does not occur when stimuli correspond to detecting similar events, but does otherwise), the proposed modeling is not a fake.

Conclusion

We propose here the use a variational framework to model and simulate, given natural image sequences, the mesoscopic collective non-standard behavior of some retinal input/output functions that correspond to the output of a subclass of the so-called K-cells. We hypothesize that from sophisticated temporal pattern recognition, to image segmentation, or specific natural statistical recognition, a unique generic two-layers non-linear filtering mechanism with feedback is implemented in the biological tissues, while not the individual but the collective behavior of the retinal cells answer for such input/output functions. Taking the retinal architecture and related biological constraints into account and considering the wider class of early-vision non-standard sensory-motor functionality known as non-standard behaviors, we use computer vision methods to propose an effective link between the observed functions and their possible implementation in the retinal network.

At the application level, we simply propose a generic visual front-end to be used for either the simulation of the sensory input of a larger simulator of the brain behavior (e.g., considering sensory-motor loops, embedded vision, ..) or for industrial applications considering such similar visual input, which concretely means degraded visual input.

Such framework raises falsifiable predictions to be verified at the biological level and points out how the retina can computationally deliver visual event candidates driving subsequent visual processes. With such framework, it is not possible to assume that the brain visual system stands on the reception of an homogeneous visual pipe of filtered images coming from the retina,

whereas it is connected to several heterogeneous sources of information at different spatial and visual scales, and different integration levels, while tuned to the natural image statistics. At the application level, on one hand, this completely changes the way artificial systems could be inspired by the biological retina. Roughly speaking, this encourages to compute in parallel several information streams depending on the sensory-motor task, instead of thinking of a large information pipe. On the other hand, this makes very questionable the idea of designing relevant visual prostheses by simply "connecting" a visual sensor array to the brain, as it is. In a nutshell, visual prostheses are likely to be though as a plural of sensors with a non-negligible data processing, before feeding the nervous system.

Acknowledgment A big thank you to Stéphane Deny for recent discussions and Pierre Kornprobst for older discussions, both at the origin of some important aspects of this work. We also acknowledge FONDECYT Nro.1120570 (Chile), CONICYT PIA Nro. 79100014 (Chile), ANR-47 CONICYT (Chile) and DGIP USM 231117.

References

Ardiles, A. O., Tapia-Rojas, C. C., Mandal, M., Alexandre, F., Kirkwood, A., Inestrosa, N. C., et al. (2012, August 06). Postsynaptic dysfunction is associated with spatial and object recognition memory loss in a natural model of Alzheimer's disease. *Proceedings of the National Academy of Sciences*.

Aubert, G., & Kornprobst, P. (2009, February). Can the nonlocal characterization of sobolev spaces by bourgain etal. be useful to solve variational problems? *SIAM Journal on Numerical Analysis*, *47*(2), 844–860.

Barlow, H. (2001, August). Redundancy reduction revisited. *Network*, *12*(3), 241–253.

Barriga-Rivera, A., & Suaning, G. J. (2011). Digital image processing for visual prosthesis: filtering implications. *Conference proceedings : ... Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Conference*, *2011*, 4860–4863.

Callaway, E. M. (1998). Local circuits in primary visual cortex of the macaque monkey. *Annual Review of Neuroscience*, *21*, 47–74.

Chambolle, A. (1995). Image segmentation by variational methods: Mumford and Shah functional and the discrete approximation. *SIAM Journal of Applied Mathematics*, *55*(3), 827–863.

Chambolle, A., & Dal Maso, G. (1999). Discrete approximations of the Mumford–Shah functional in dimension two. *M2AN*, *33*(4), 651–672. ((also available as Technical report 9820 from Université Paris Dauphine, Ceremade))

Ciocchi, S., Herry, C., Grenier, F., Wolff, S. B., Letzkus, J. J., Vlachos, I., et al. (2010, November 11). Encoding of conditioned fear in central amygdala inhibitory circuits. *Nature*, *468*(7321), 277–282.

Cowey, A. (2010, September 14). Visual System: How Does Blindsight Arise? *Curr Biol*, *20*(17), R702–R704.

Dan, Y., Atick, J. J., & Reid, R. C. (1996, May 15). Efficient coding of natural scenes in the lateral geniculate nucleus: experimental test of a computational theory. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, *16*(10), 3351–3362.

Delgado, L. M., Vielma, A. H., Kähne, T., Palacios, A. G., & Schmachtenberg, O. (2009, June 10). The GABAergic system in the retina of neonate and adult Octodon degus, studied by immunohistochemistry and electroretinography. *The Journal of comparative neurology*, *514*(5), 459–472.

Dong, D. W. (2001). Spatiotemporal Inseparability of Natural Images and Visual Sensitivities

Motion Vision. In J. M. Zanker & J. Zeil (Eds.), *Motion vision* (pp. 371–380). Berlin, Heidelberg: Springer Berlin Heidelberg.

Field, G., Sher, A., Gauthier, J., Greschner, M., Shlens, J., Litke, A., et al. (2007, November). *The Journal of Neuroscience*, *27*(48), 13261-13272.

Field, G. D., Sher, A., Gauthier, J. L., Greschner, M., Shlens, J., Litke, A. M., et al. (2007, November 28). Spatial properties and functional organization of small bistratified ganglion cells in primate retina. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, *27*(48), 13261–13272.

Gauthier, J. L., Field, G. D., Sher, A., Greschner, M., Shlens, J., Litke, A. M., et al. (2009, April 7). Receptive fields in primate retina are coordinated to sample visual space more uniformly. *PLoS biology*, *7*(4), e63+.

Gérard, L., Kornprobst, P., & Viéville, T. (2007). From variational to spiking network image-segmentation techniques. In *Perception 36 ecvp abstract supplement.*

Gollisch, T., & Meister, M. (2008). Rapid neural coding in the retina with relative spike latencies. *Science*, *319*, 1108–1111. (DOI: 10.1126/science.1149639)

Gollisch, T., & Meister, M. (2010, January 28). Eye Smarter than Scientists Believed: Neural Computations in Circuits of the Retina. *Neuron*, *65*(2), 150–164.

Greschner, M., Shlens, J., Bakolitsa, C., Field, G. D., Gauthier, J. L., Jepson, L. H., et al. (n.d.). Correlated firing among major ganglion cell types in primate retina.

Hendry, S. H., & Reid, R. C. (2000). The Koniocellular Pathway in Primate Vision. *Annual Review of Neuroscience*, *23*(1), 127–153.

Hochstein, S., & Shapley, R. M. (1976). Linear and nonlinear spatial subunits in Y cat retinal ganglion cells. *J Physiol*, *262*, 265–284.

Hornik, K., Stinchcombe, M., & White, H. (1989). Multilayer feedforward networks are universal approximators. *Neural Networks*, *2*, 359–366.

Hosoya, T., Baccus, S. A., & Meister, M. (2005, July 7). Dynamic predictive coding by the retina. *Nature*, *436*(7047), 71–77.

Hyvärinen. (2009). *Natural Image Statistics.* Springer-Verlag.

Karklin, Y., & Lewicki, M. S. (2006). Is Early Vision Optimized for Extracting Higher-order Dependencies? In Y. Weiss, B. Schölkopf, & J. Platt (Eds.), *Advances in neural information processing systems 18.* Cambridge, MA: MIT Press.

Kayser, C., Salazar, R. F., & Konig, P. (2003, September). Responses to natural scenes in cat V1. *Journal of neurophysiology*, *90*(3), 1910–1920.

Keskes, M. (2011). *Modélisation des premières étapes de la vision avec des réseaux de neurones.* Unpublished master's thesis.

Koivisto, M., Railo, H., Revonsuo, A., Vanni, S., & Salminen-Vaparanta, N. (2011, February 16). Recurrent Processing in V1/V2 Contributes to Categorization of Natural Scenes. *The Journal of Neuroscience*, *31*(7), 2488–2492.

Lamme, V., & Roelfsema, P. R. (2000, November 1). The distinct modes of vision offered by feedforward and recurrent processing. *Trends in Neurosciences*, *23*(11), 571–579.

Lamme, V. A. (2001, April). Blindsight: the role of feedforward and feedback corticocortical connections. *Acta psychologica*, *107*(1-3), 209–228.

Li, Z. (1992). Different retinal ganglion cells have different functional goals. *International Journal of Neural Systems*, *3*, 237–248.

Litke, A. M., Bezayiff, N., Chichilnisky, E. J., Cunningham, W., Dabrowski, W., Grillo, A. A., et al. (2004, August). What Does the Eye Tell the Brain?: Development of a System for the Large-Scale Recording of Retinal Output Activity. *Nuclear Science, IEEE Transactions on*, *51*, 1434–1440.

Masland, R. H., & Martin, P. R. (2007, August 07). The unsolved mystery of vision. *Current Biology*, *17*(15), R577–R582.

Mumford, D., & Shah, J. (1985, June). Boundary detection by minimizing functionals. In *Proceedings of the international conference on computer vision and pattern recognition* (pp. 22–26). San Francisco, CA: IEEE.

Nassi, J. J., & Callaway, E. M. (2009). Parallel processing strategies of the primate visual system. *Nat. Rev. Neurosci.*, *10*(5), 360–372.

Olveczky, B. P., Baccus, S. A., & Meister, M. (2003, May 22). Segregation of object and background motion in the retina. *Nature*, *423*(6938), 401–408.

Peichl, L., Chavez, A. E., Ocampo, A., Mena, W., Bozinovic, F., & Palacios, A. G. (2005, June 6). Eye and vision in the subterranean rodent cururo (Spalacopus cyanus, Octodontidae). *The Journal of comparative neurology*, *486*(3), 197–208.

Pitkow, X., & Meister, M. (2012, April 11). Decorrelation and efficient coding by retinal ganglion cells. *Nat Neurosci*, *15*(4), 628–635.

Rullen, R. V., & Thorpe, S. (2001). Rate coding versus temporal order coding: What the retina ganglion cells tell the visual cortex. *Neural Computing*, *13*(6), 1255-1283.

Schwartz, G., & Michael. (2008, April 01). Sophisticated Temporal Pattern Recognition in Retinal Ganglion Cells. *Journal of Neurophysiology*, *99*(4), 1787–1798.

Shlens, J., Rieke, F., & Chichilnisky, E. (2008, October 27). Synchronized firing in the retina. *Current Opinion in Neurobiology*.

Simoncelli, E. P. (2003, April). Vision and the statistics of the visual environment. *Current Opinion in Neurobiology*, *13*(2), 144–149.

Simoncelli, E. P., & Olshausen, B. A. (2001). Natural Image Statistics and Neural Representation. *Annual Review of Neuroscience*, *24*(1), 1193–1216.

Simoncini, C., Perrinet, L., Montagnini, A., Mamassian, P., & Masson, G. (2012). *Nature Neuroscience*, *Epub ahead of print*.

Soo, F. S., Schwartz, G. W., Sadeghi, K., & Berry, M. J. (2011, February 9). Fine Spatial Information Represented in a Population of Retinal Ganglion Cells. *The Journal of Neuroscience*, *31*(6), 2145–2155.

Sterling, P. (2004). How retinal circuits optimize the transfer of visual information. In L. M. Chalupa & J. S. Werner (Eds.), *The visual neurosciences* (pp. 234–259). MIT Press, Cambridge MA.

Subbarao, M. (1990, June). Bounds on time-to-collision and rotational component from first-order derivatives of image flow. *Computer Vision, Graphics, And Image Processing*, *52*(3), 329–341.

Teftef, E. (2012). *Formalisation de la transformation analogique / événementielle des mécanismes non-standards des cellules ganglionnaires de la rétine.* Unpublished master's thesis, Université de Paris 6 et Université El Manar de Tunis.

Teftef, E., & Viéville, T. (2012). Formalization of the input/output retinal transformation regarding non-standard ganglion cells behavior. In *The neuromp'12 keops workshop.* (submitted)

Thorpe, S., & Fabre-Thorpe, M. (2001). Seeking categories in the brain. *Science*, *291*, 260–263.

Thorpe, S., Fize, D., & Marlot, C. (1996, June 06). Speed of processing in the human visual system. *Nature*, *381*(6582), 520–522.

Tkačik, G., Ghosh, A., Schneidman, E., & Segev, R. (2012, January 17). Retinal adaptation and invariance to changes in higher-order stimulus statistics.

Torralba, A., & Oliva, A. (2003). Statistics of natural image categories. In *Network: Computation in neural systems* (Vol. 14, pp. 391–412).

Truccolo, W. (2001, June). Dynamic temporal decorrelation: An information-theoretic and

biophysical model of the functional role of the lateral geniculate nucleus. *Neurocomputing*, *38-40*(1-4), 993–1001.

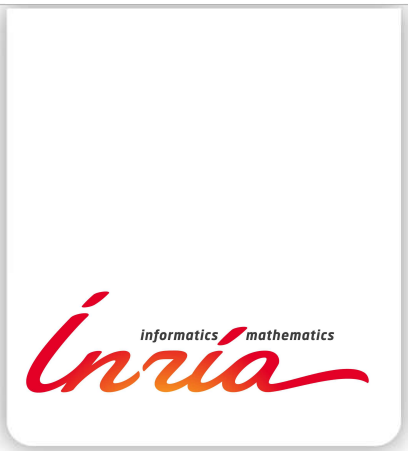
Viéville, T., Chemla, S., & Kornprobst, P. (2007). How do high-level specifications of the brain relate to variational approaches? *J. Physiol. Paris*, *101*.

Viéville, T., & Crahay, S. (2004). Using an hebbian learning rule for multi-class svm classifiers. *Journal of Computational Neuroscience*, *17*(3), 271–287.

Werblin, F. S. (2011, August 1). The retinal hypercircuit: a repeating synaptic interactive motif underlying visual function. *The Journal of physiology*, *589*(Pt 15), 3691–3702.

Wohrer, A. (2008). *Model and large-scale simulator of a biological retina with contrast gain control*. Unpublished doctoral dissertation, University of Nice Sophia-Antipolis.

Yoonessi, A., & Yoonessi, A. (2011, April). Functional assessment of magno, parvo and koniocellular pathways; current state and future clinical applications. *Journal of ophthalmic & vision research*, *6*(2), 119–126.



RESEARCH CENTRE
BORDEAUX – SUD-OUEST

351, Cours de la Libération
Bâtiment A 29
33405 Talence Cedex

Publisher
Inria
Domaine de Voluceau - Rocquencourt
BP 105 - 78153 Le Chesnay Cedex
inria.fr

ISSN 0249-6399