Visual gaze selection modeling from visual input

Andrea Carbone
Dept. Computer and System Sciences, Sapienza
Email: carbone@dis.uniroma1.it

Fiora Pirri
Dept. Computer and System Sciences, Sapienza
Email: pirri@dis.uniroma1.it

***Abstract*— In this work we present a biologically motivated framework for the modeling of the visual scene exploration preference. We aim at capturing the statistical patterns that are elicited by the subjective visual selection and reproduce them via a computational system.**

I. Introduction

The visual exploration of the world, the *scan-path*, is performed by human beings in a very efficient way by programming a sequence of *saccades* on the visual scene in order to project the selected spatial focus of attention onto the higher resolution surface of the retina. A crucial aspect that has gained lot of interest is how the early visual system encodes and processes the visual stimuli reaching the retina [2]. Lot of effort has been devoted to the discovery of the sensory coding mechanism. Early works on this subject highlighted the importance in modeling the statistical regularities of the visual input, following the assumption that the visual system exploits those regularities to efficiently code the visual information [1]. Nonetheless as argued in several works [3], the purpose of the early visual processing is to produce a sparsified representation of the visual input rather than a compression (in the minimum code length interpretation) [8]. A very simple but successful model assumes that any given set of natural images (or patches) $I(x, y)$ can be generated by a linear combination of features W or basis vectors B:

$$I(x,y) = \sum_{i=1}^{n} B_i(x,y)s_i \quad , \quad s_i = \sum_{x,y} W_i(x,y)\ I(x,y) \quad (1)$$

where the coefficients s_i are the coefficients weighting the i-th basis vector.

A sparse code is a code whose response distribution is usually sharply peaked and long tailed. The idea supporting sparseness is therefore that only a small subset of a large population of cells will be active when presented to a specific family of visual stimuli. In this work we focus exclusively on the characterisation on the sensory model of the simple and complex cells at the early stages of cortical processing (V1) which exhibit a localised, oriented and bandpass behaviour [4]. In this work we are using a standard dataset comprising 101 images depicting urban and natural scenes with corresponding eye-tracked data from 31 subjects. Each image has associated an individual (per subject) and a cumulative fixation distance-map, for further reference see [6].

II. Our Approach

We suggest a workflow combining several aspects of computational neuroscience and machine learning which aim at identifying those statistical patterns which trigger a subject's spot of attention towards specific locations.

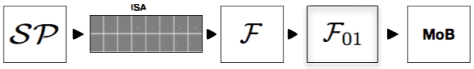

Fig. 1. From the scanpath to the Mixture of Bernoulli. Projection to the ISA feature space, binarization and finally EM estimation of the mixture parameters.

A computational model is derived by firstly estimating a mixture of multivariate Bernoulli variables out of a binary activation map representing the visual receptive field responses to a target sequence of fixation patches, Fig. 1. Afterwards, each component of the mixture is taken as modeling a specific family of receptive field responses which (statistically) share a common pattern of activation. A dichotomized Gaussian representation lets us select for each component the most representative sample that we elect as prototypal example of the population. The set of prototypal samples are related to a binary representation which bring us to a natural interpretation of the patterns as a collection of *on* and *off* maps. The on-off channels map respectively those channels whose response are expected to lie on the tail (a relatively high value) or the peak (close to zero), Fig. 2.

A. Receptive field estimation

The Independent Subspace Analysis (ISA) [5] is based on the maximisation of independence between linear subspaces. ISA is configured as a first linear filtering followed by an energy pooling stage. The result is a set of linear filters whose pooled response reproduce closely the response properties of complex cells in V1 (independence of the response magnitude w.r.t. the phase of the signal). The output response u_j of the j-th subspace S_j is computed as follows:

$$u_j = \sqrt{\sum_{i \in S_j} s_i^2} \quad (2)$$

where the s_i are computed as in Eq. 1.

The ISA basis is computed on a set of 60000 (24×24) pixel image patches randomly selected from the image collection.

B. Scan-path projection on ISA basis and Mixture of Bernoulli estimation

We define the *target* scanpath $\mathcal{SP}$ as a set of fixation patches filtered out from a gaze-tracked sequence. The matrix $\mathcal{F}$ contains on each row the u_j coefficients computed as in

Fig. 2. From the MoB to the final linear filtering cascade.

Eq. 2. The size of $\mathcal{F}$ depends on the number N of fixations and on the geometry of the used ISA basis. In our experiments we computed a complete linear basis of dimension 576 for a 24×24 image space. The subspace size was set to 8, thus the pooled feature space has dimension 72. The $\mathcal{F} \in \mathcal{R}^{N \times 72}$ is therefore our input dataset. Our goal is to capture the general pattern of activation of the single subspace feature channels, therefore we binarize each scan-path feature vector setting to 1 all the values which are above the standard deviation of the overall values (w.r.t. the single vector) and to 0 the values below. We obtain $\mathcal{F}_{01}$ a binary matrix of the same size as $\mathcal{F}$ which is the spike train activation map corresponding to the original train of scan-path patches. We model the $\mathcal{F}_{01}$ as samples of a mixture of multivariate Bernoulli (MoB).The parameters of the mixture have been estimated via expectation-maximisation (EM) algorithm for Bernoulli mixtures Fig. 1.

III. The Computational Model

We link the binary spike population estimated mixture to its corresponding dichotomized Gaussian (DG) [7] generative counterpart. The properties of such decomposition are exploited by finding out the most probable subspace (in the DG space). We show how such subspace encodes a representative activity pattern of the receptive field responses elicited by the target visual input.

The DG distribution is a statistical tool described in [7] used to generate spike trains with given first and second order statistics by considering dicothomized samples of a multivariate Gaussian distribution $\mathcal{N}(\gamma, \Lambda)$. For each of the M components the set $\mathcal{F}_{01}^m \in \mathcal{F}_{01}$ is the subset of the binarized scan-path samples that are more likely to belong to the m-th component . From each $\mathcal{F}_{01}^m$ we estimate the corresponding DG model $\mathcal{N}(\gamma_m, \Lambda_m)$. At this point, we have at our disposal a generative model of the binarized input dataset representing the full scan-path. It inherits the structure from the previously estimated MoB, resulting in a weighted sum of M multivariate DG distributions. The idea is to make use of the MoB and its DG model decomposition to implement a filtering cascade that will output a high saliency value corresponding to local images patches that have a receptive field pattern of responses which is close to the modeled scan-path. Intuitively, in the dicothomized representation we can identify the most probable binary pattern that can occur by considering the sign of γ_m. In fact, in a N dimensional space there are 2^N subspaces centered on the origin depending on the sign chosen on the n-th individual dimension. Each of these subspaces by definition is mapped to a specific binary pattern.We want to identify the most probable binary pattern that can be sampled from $\mathcal{N}(\gamma_m, \Lambda_m)$. The subspace where the mean of the DG falls it is of course the subspace whose cumulative distribution is greater than the others. Therefore we take $sign(\gamma_m)$ as the prototypical binary activation pattern which approximates the m-th component. The negative sign will be mapped to a 0 and the positive to a 1. The ISA-subspace feature channels corresponding to the ones and the zeroes will model respectively the *on* and *off* linear filtering cascades for each DG component. The on and off responses to a single patch input are given by:

$$on = \sum_{i=1}^{M} \pi_i \prod_{j=1}^{D} u_j b_{ij} \quad , \quad off = \sum_{i=1}^{M} \pi_i \prod_{j=1}^{D} u_j (1 - b_{ij}) \quad (3)$$

where π_i are the mixture coefficients of the i-th component, u_j are the subspace coefficients as in Eq. 2 for the input patch and the b_{ij} are 1 or 0 depending on the previous interpretation of the mean of the DG model of the i-th component. The on-off maps computed on every (24×24) patch of the input image are then merged together to get the final saliency map $\mathcal{SM}$ Fig. 2. In Fig. 3 a test image from the category "buildings" of the dataset, the saliency map and the fixation distance-map representing the eye-tracking data of all the 31 subjects. The brighter regions (red in the color version) indicate high saliency values.

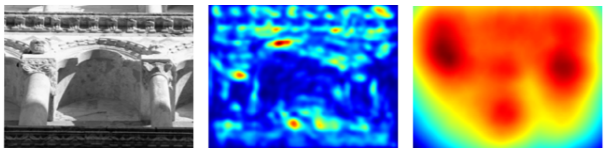

Fig. 3. An input test image, its computed saliency map and the reference fixation distance map.

Acknowledgment

The work was supported by the EC FP7 IST project NIFTi-247870.

References

[1] H. Barlow, "The exploitation of regularities in the environment by the brain." *Behav Brain Sci*, vol. 24, no. 4, pp. 602–7; discussion 652–71, Aug 2001.

[2] M. Carandini, J. Demb, V. Mante, D. Tolhurst, Y. Dan, B. A. Olshausen, J. Gallant, and N. Rust, "Do we know what the early visual system does?" *J. of Neuroscience*, vol. 25, no. 46, p. 10577, 2005.

[3] D. J. Field, "What is the goal of sensory coding?" *Neural Comp*, vol. 6, pp. 559–601, Jan 1994.

[4] D. Hubel and T. Wiesel, "Receptive fields and functional architecture of monkey striate cortex." *The Journal of Physiology*, vol. 195, pp. 215–243, Jan 1968.

[5] A. Hyvarinen and U. Koster, "Complex cell pooling and the statistics of natural images." *Network: Comp in Neur Sys*, vol. 18, no. 2, pp. 81–100, 2007.

[6] G. Kootstra, A. Nederveen, and B. de Boer, "Paying attention to symmetry." *Proc of BMVC*, pp. 1115–1125, 2008.

[7] J. Macke, P. Berens, A. Ecker, A. Tolias, and M. Bethge, "Generating spike trains with specified correlation coefficients." *Neur Comp*, vol. 21, no. 2, pp. 397–423, 2009.

[8] B. A. Olshausen and D. J. Field, "Natural image statistics and efficient coding." *Comp in Neural Sys*, vol. 7, no. 2, pp. 333–339, 1996.