Vision Research 50 (2010) 2295–2307

Contents lists available at ScienceDirect

Vision Research

journal homepage: www.elsevier.com/locate/visres

Biologically plausible saliency mechanisms improve feedforward object recognition

Sunhyoung Han *, Nuno Vasconcelos

Electrical and Computer Engineering, University of California, San Diego, 9500 Gilman Drive, La Jolla, CA 92093-0407, United States

ARTICLE INFO

Article history:
Received 16 October 2009
Received in revised form 13 February 2010

Keywords:
Discriminant saliency
HMAX
Object recognition
Visual attention
Biologically plausible recognition
Neural models

ABSTRACT

The biological plausibility of statistical inference and learning, tuned to the statistics of natural images, is investigated. It is shown that a rich family of statistical decision rules, confidence measures, and risk estimates, can be implemented with the computations attributed to the standard neurophysiological model of V1. In particular, different statistical quantities can be computed through simple re-arrangement of lateral divisive connections, non-linearities, and pooling. It is then shown that a number of proposals for the measurement of visual saliency can be implemented in a biologically plausible manner, through such re-arrangements. This enables the implementation of biologically plausible feedforward object recognition networks that include explicit saliency models. The potential of combined attention and recognition is illustrated by replacing the first layer of the HMAX architecture with a saliency network. Various saliency measures are compared, to investigate whether (1) saliency can substantially benefit visual recognition and (2) the benefits depend on the specific saliency mechanisms implemented. Experimental evaluation shows that saliency does indeed enhance recognition, but the gains are not independent of the saliency mechanisms. Best results are obtained with top-down mechanisms that equate saliency to classification confidence.

1. Introduction

The effectiveness and speed of biological solutions to the object recognition problem have long been a source of inspiration for recognition algorithms. The introduction of the back-propagation algorithm (Rumelhart, Smolenksy, Mcclelland, & Hinton, 1986) established a framework for the automated design of recognition networks, and was highly successful for a number of problems. In particular, convolutional networks were shown to be highly competitive with the best non-biological classifiers for tasks such as hand-written character recognition (Lecun, Bottou, Bengio, & Haffiner, 1998). More recent results, by Thorpe, Fize, and Marlot (1996), on the ability of human subjects to categorize natural scenes, showed that such tasks can be performed with high accuracy (close to 94%) and very quickly (in less than 150 ms). The fact that such low recognition times leave no room for propagation of feedback across cortical areas, reinforced the significance of feedforward networks in visual recognition, at least in its early stages. It also spurred a renewed interest in the family of feedforward architectures, of which the most recent popular element is the HMAX network of Riesenhuber and Poggio (1999) and Serre et al. (2007). This network emulates the organization of the visual system as a cascade of layers of simple and complex cells (Hubel & Wiesel, 1962), and has been recently shown to achieve state-of-the-art performance for a number of recognition tasks (Mutch & Lowe, 2008).

There are, however, two important limitations of the HMAX model. First, because the organization of the network lacks a clear *computational justification*, HMAX networks also lack a principled *optimality criterion* and *training* algorithm. This limits their relevance as an explanation for the underlying biological computations. Second, HMAX networks do not account for the psychophysical evidence on the important role played by *visual attention* in top-down processes such as object recognition (Yarbus, 1967). This limitation has been somewhat mitigated by research on recognition within multi-object displays, which complements the HMAX network with serial attention mechanisms (Miau, Papageorgiou, & Itti, 2001; Walther & Koch, 2006). In these methods, saliency is computed with an independent bottom-up network, which (1) acts as a "front-end" to the HMAX network, selecting patches of the visual field to recognize (Miau et al., 2001) or (2) modulates the connections of some HMAX units, serially directing attention to different proto-objects in the field of view (Walther & Koch, 2006). None of these works can account for the role of top-down attention in recognition, or the benefits of saliency in single object displays. These benefits have been documented in the computer vision literature (Kadir & Brady, 2001; Mikolajczyk & Schmid, 2004; Sebe & Lew, 2003), but with recourse to interest-point detectors that are not biologically plausible. Within the HMAX literature, it has been shown that limiting the spatial pooling performed by some of the HMAX units can lead to non-trivial recognition

* Corresponding author. Fax: +1 858 534 1483.
E-mail address: s1han@ucsd.edu (S. Han).

0042-6989/$ - see front matter © 2010 Elsevier Ltd. All rights reserved.
doi:10.1016/j.visres.2010.05.034

improvements (Mutch & Lowe, 2008). This, however, has been done in a somewhat ad-hoc form, by restricting the receptive fields of these units to a pre-defined window size. To the best of our knowledge, no *formal* connection has been established between HMAX itself and visual attention.

In this work, we suggest a modification of the HMAX architecture that makes the connection between recognition and visual saliency explicit. We start by investigating the *biological plausibility* of statistical inference and learning tuned to the statistics of natural images. Building on prior work by Gao and Vasconcelos (2009), we show that a rich family of statistical *decision rules, confidence measures, and risk estimates,* can be implemented with the computations attributed to the standard neurophysiological model of V1 (Carandini, Heeger, & Movshon, 1997; Carandini et al., 2005; Heeger, 1992; Hubel & Wiesel, 1962): a combination of linear filtering, divisive normalization, non-linearities, and spatial pooling. In fact, it is shown that all these computations have precise *statistical meaning*, contributing to an overall probabilistic interpretation where simple cells compute posterior probabilities and complex cells estimate statistical risks. It follows that *a number of statistical operators can be implemented with biological hardware*, through simple re-arrangement of lateral divisive connections, non-linearities, and pooling. We next establish a connection to *saliency* mechanisms, by showing that various proposals for the measurement of visual saliency, from both the biological and computer vision literatures, *can be implemented* with biologically plausible reconfigurations of the standard neurophysiological model. By replacing the first layer of the HMAX architecture with these saliency networks, we conduct a rigorous experimental study of three questions at the intersection of attention and feedforward object recognition: (1) whether saliency benefits visual recognition, (2) whether the gains depend on the type of saliency considered (e.g. top-down vs. bottom-up) or even the specific saliency algorithms, and (3) whether max-based pooling has an advantage over the classical linear operator. We note that the goal is not to investigate whether saliency is beneficial as a means to serialize recognition when there are multiple objects within the field of view, as has been done in Miau et al. (2001), Walther and Koch (2006), or whether there are gains in complementing recognition with an independent saliency path. Instead, we consider the question of whether saliency is *intrinsically* important for recognition, even when there is a single object in the field of view, as is suggested by computer vision research. Or, in other words, whether in addition to its predominant role within the "where" pathway, saliency also plays a role within the "what" pathway of object recognition. It is shown that the addition of saliency can *significantly improve* recognition performance, but that this is not independent of the saliency principle adopted. Best results are obtained with top-down saliency mechanisms that equate saliency to classification confidence.

2. Method

We study the biological plausibility of statistical inference tuned to the statistics of natural images. We start by reviewing some known properties of these statistics, then consider statistical inference, and finally the learning problem.

2.1. Natural image statistics

Various authors have shown that the empirical distribution of the response X of a band-pass filter to a wide variety of natural imagery is accurately modeled by the generalized Gaussian distribution (GGD) (Buccigrossi & Simoncelli, 1999; Do & Vetterli, 2002; Huang & Mumford, 1999). This distribution is defined as

$$P_X(x;\alpha,\beta) = \frac{\beta}{2\alpha\Gamma(1/\beta)} e^{-\left(\frac{|x|}{\alpha}\right)^{\beta}} \quad (1)$$

where $\Gamma(z) = \int_0^\infty e^{-t}t^{z-1}\,dt$, $t>0$ is the Gamma function, α a *scale* parameter, and β a parameter that controls the *shape* of the distribution.

The parameters α, β can be learned in multiple ways, including the method of moments (Huang & Mumford, 1999), maximum likelihood (Do & Vetterli, 2002), or Bayesian maximum a posteriori (MAP) estimation (Gao & Vasconcelos, 2009). We adopt the latter, using a (Gamma distributed) conjugate prior for the scale parameter α. Given a sample of training observations $\mathcal{D} = \{x_1,\ldots,x_n\}$, this leads to Gao and Vasconcelos (2009)

$$\hat{\alpha}^{\beta}_{MAP} = \frac{1}{\kappa}\left(\sum_{j=1}^{n}|x_j|^{\beta} + v\right), \quad \text{with} \quad \kappa = \frac{n+\eta}{\beta} \quad (2)$$

where η and v are prior hyper-parameters. The details of the prior are not crucially important, as its role is simply to regularize the feature responses, so as to prevent a null scale estimate. In our implementation we use $\eta = 1$ and $v = 10^{-3}$. The MAP estimate of the shape parameter β is more complex. However, for natural images this parameter tends to be fairly stable, usually taking values between .5 and .8 (Srivastava, Lee, Simoncelli, & Zhu, 2003). We have found $\beta = .5$ to maximize the likelihood of a large sample of responses of a set of Gabor filters to a random collection of natural images. This is illustrated in Fig. 1, which shows the log–probability histogram of the Gabor responses and the MAP GGD fit for $\beta = .5$. This value was used in all experiments reported in this work.

2.2. Statistical inference

The biological plausibility of probabilistic inference with GGD stimuli was studied in Gao and Vasconcelos (2009). This work has shown that, for such stimuli, the fundamental computations of probabilistic inference and learning can be implemented with the standard computational model of simple and complex cells (Carandini et al., 1997, 2005; Heeger, 1992; Hubel & Wiesel, 1962). In what follows, we extend the procedures introduced by Gao and Vasconcelos (2009) to show that a much broader set of computations, summarized in Table 1, is biologically plausible. These computations are described in the second column of the table. Although their biological implementation turns out to be possible with subtle modifications to the computations of Gao and Vasconcelos (2009), namely the introduction of various non-linear-

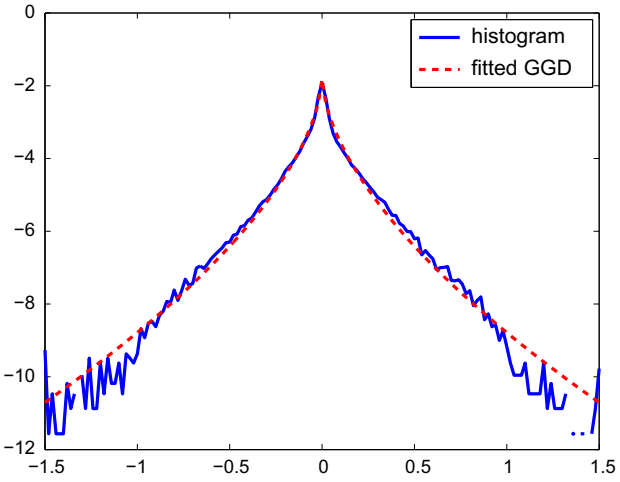

Fig. 1. Histogram of responses of a set of Gabor filters to a collection of natural images, and its MAP fit by the GGD model with $\beta = .5$.

Table 1
Operations of statistical inference under the GGD model. $\psi(x)$ is defined as $\psi(x) = \frac{1}{2}\log\frac{x}{1-x}$

Operation	Definition	Under GGD statistics	Notes
Single observation x inference			
Neg. log-likelihood (NLL)	$-\log P_X(x)$	$l^\beta_\alpha(x) = \left(\frac{\lvert x\rvert}{\alpha}\right)^\beta + K$	$K = \log\frac{2\alpha\Gamma(1/\beta)}{\beta}$
Log-likelihood ratio (LLR)	$\log\frac{P_{X\lvert Y}(x\lvert 1)}{P_{X\lvert Y}(x\lvert 0)}$	$g(x) = l^\beta_{\alpha_0}(x) - l^\beta_{\alpha_1}(x)$	
Target posterior (TP)	$P_{Y\lvert X}(1\lvert x)$	$\sigma[g(x)]$	$\sigma(x) = (1+e^{-x})^{-1}$
Information	$I(Y; X = x)$	$\xi\{\sigma[g(x)]\}$	$\xi(x) = \log 2 + x\log x + (1-x)\log(1-x)$
Measures of detection confidence			
LLRC(x)	$\tilde{\psi}\{P_{Y\lvert X}(1\lvert x)\}$	$\tilde{\psi}\{\sigma[g(x)]\}$	$\tilde{\psi}(x) = \begin{cases}\psi(x), & x \geq .5\\ 0, & x < .5\end{cases}$
IC(x)	$\tilde{\xi}\{P_{Y\lvert X}(1\lvert x)\}$	$\tilde{\xi}\{\sigma[g(x)]\}$	$\tilde{\xi}(x) = \begin{cases}\xi(x), & x \geq .5\\ 0, & x < .5\end{cases}$
Empirical risks based on sample $\mathscr{R} = \{x_1, \ldots, x_n\}$			
Expected NLL	$E_X[-\log P_X(x)]$	$\frac{1}{n}\sum_{i=1}^n \alpha,\beta(x_i)$	$H[X]$
Expected LLR	$E_X\left[\log\frac{P_{X\lvert Y}(x\lvert 1)}{P_{X\lvert Y}(x\lvert 0)}\right]$	$\frac{1}{n}\sum_{i=1}^n g(x_i)$	$KL[P_X(x)\lVert P_{X\lvert Y}(x\lvert 0)] - KL[P_X(x)\lVert P_{X\lvert Y}(x\lvert 1)]$
MI	$E_X[I(Y; X = x)]$	$\frac{1}{n}\sum_{i=1}^n \xi\{\sigma[g(x_i)]\}$	$I(Y; X)$
Expected confidence (LLR)	$E_X[\mathrm{LLR}(x)]$	$\frac{1}{n}\sum_{i=1}^n \tilde{\psi}\{\sigma[g(x_i)]\}$	$KL[P_{X\lvert y}(x\lvert 1)\lVert P_{X\lvert Y}(x\lvert 0)]$
Expected confidence (MI)	$E_X[\mathrm{IC}(x)]$	$\frac{1}{n}\sum_{i=1}^n \tilde{\xi}\{\sigma[g(x_i)]\}$	

ities, this extension substantially broadens the scope of the underlying computational framework. For example, the operations now considered are critical to the design of networks that address top-down problems such as object recognition. In fact, as will be shown in Section 3.3, the performance of such top-down networks can be quite sensitive to the precise choice of statistical inference principle, and associated non-linearities.

Table 1 is organized in three sections. The first reports to inference from a single observation x. It starts with the most atomic computation of statistical inference: the evaluation of the log–probability $\log P_X(x)$ of an observation x. A perceptual system can use this probability to make optimal decisions regarding the classification of x with respect to a target and a null hypothesis. These are identified by a class label Y that takes the values $Y = 1$ for the target and $Y = 0$ for the null hypothesis. Optimal decision-making is frequently defined in the minimum probability of error (MPE) sense, under which the optimal procedure is the Bayes decision rule (Duda, Hart, & Stork, 2001). This consists of thresholding the log-likelihood ratio (LLR)

$$\log\frac{P_{X|Y}(x|1)}{P_{X|Y}(x|0)} \tag{3}$$

and selecting the target hypothesis whenever this ratio is above threshold. An equivalent implementation of this decision rule is to choose the target hypothesis when the posterior target probability, $P_{Y|X}(1|x)$, is above 1/2. The process is illustrated in Fig. 2, for an object recognition problem where the target is the class of airplanes. Given a set of example images from this class, and a set of examples from the null hypothesis (in this case any object other than a plane), the visual system relies on a set of bandpass (e.g. Gabor) filters to extract visual features characteristic of the two classes. The GGDs that best fit the distributions, $P_{X|Y}(x|i), i \in \{0,1\}$, of filter responses under the two hypotheses are then estimated. Given a new image, the corresponding features are extracted, and the LLR of (3) is computed, using these GGDs. Thresholding this quantity then produces a binary map that indicates the locations of the target within the visual field.

The LLR is one of various quantities that play an important role in statistical inference and optimal decision making. Table 1 includes a number of others, which we review in more detail in the remainder of this section. A graphical illustration of these measures, in the context of object recognition, is presented in Figs. 4 and 5. An alternative optimality criteria for decision making, commonly referred as infomax (Linsker, 1988), is to maximize the information about the class label Y. This criterion underlies many classification procedures proposed in the machine learning literature, including logistic regression and some forms of boosting (logitBoost) (Friedman, Hastie, & Tibshirani, 2000; Hastie, Tibshirani, & Friedman, 2001). Its maximization has also been proposed as a fundamental principle for the organization of perceptual systems (Barlow, 2001; Linsker, 1988). In this case, inference is based on the information

$$I(Y; X = x) = \sum_i P_{Y|X}(i|x)\log\frac{P_{X,Y}(x,i)}{P_X(x)P_Y(i)} \tag{4}$$

that the observation x provides about the class Y.

The second section of Table 1 refers to the evaluation of confidence measures. These complement the decision that x belongs to the target class (target detection), by quantifying how confident the classifier is about this decision. Obviously, the confidence measure should be derived from the principle used for inference. This leads to two confidence measures based on the likelihood ratio

$$\mathrm{LLRC}(x) = \begin{cases}\frac{1}{2}\log\frac{P_{Y|X}(1|x)}{P_{Y|X}(0|x)}, & P_{Y|X}(1|x) \geq .5\\ 0 & P_{Y|X}(1|x) < .5,\end{cases} \tag{5}$$

and the information measure

$$\mathrm{IC}(x) = \begin{cases}I(Y; X = x), & P_{Y|X}(1|x) \geq .5\\ 0 & P_{Y|X}(1|x) < .5.\end{cases} \tag{6}$$

An important property is that, in both cases, the confidence measure is "one-sided", i.e. non-zero only if x is classified as a target. Although undesirable for the bottom-up problems considered in Gao and Vasconcelos (2009), we will see that this property becomes quite important for success in top-down problems, such as recognition. The two measures can be expressed as a transformation of the posterior target probability $P_{Y|X}(1|x)$. As indicated in Table 1, these transformations are

$$\tilde{\psi}(x) = \begin{cases}\frac{1}{2}\log\frac{x}{1-x}, & x \geq .5\\ 0, & x < .5\end{cases} \tag{7}$$

for LLRC(x) and

$$\tilde{\xi}(x) = \begin{cases}\log 2 + x\log x + (1-x)\log(1-x), & x \geq .5\\ 0, & x < .5\end{cases} \tag{8}$$

for IC(x). They are shown in Fig. 3.

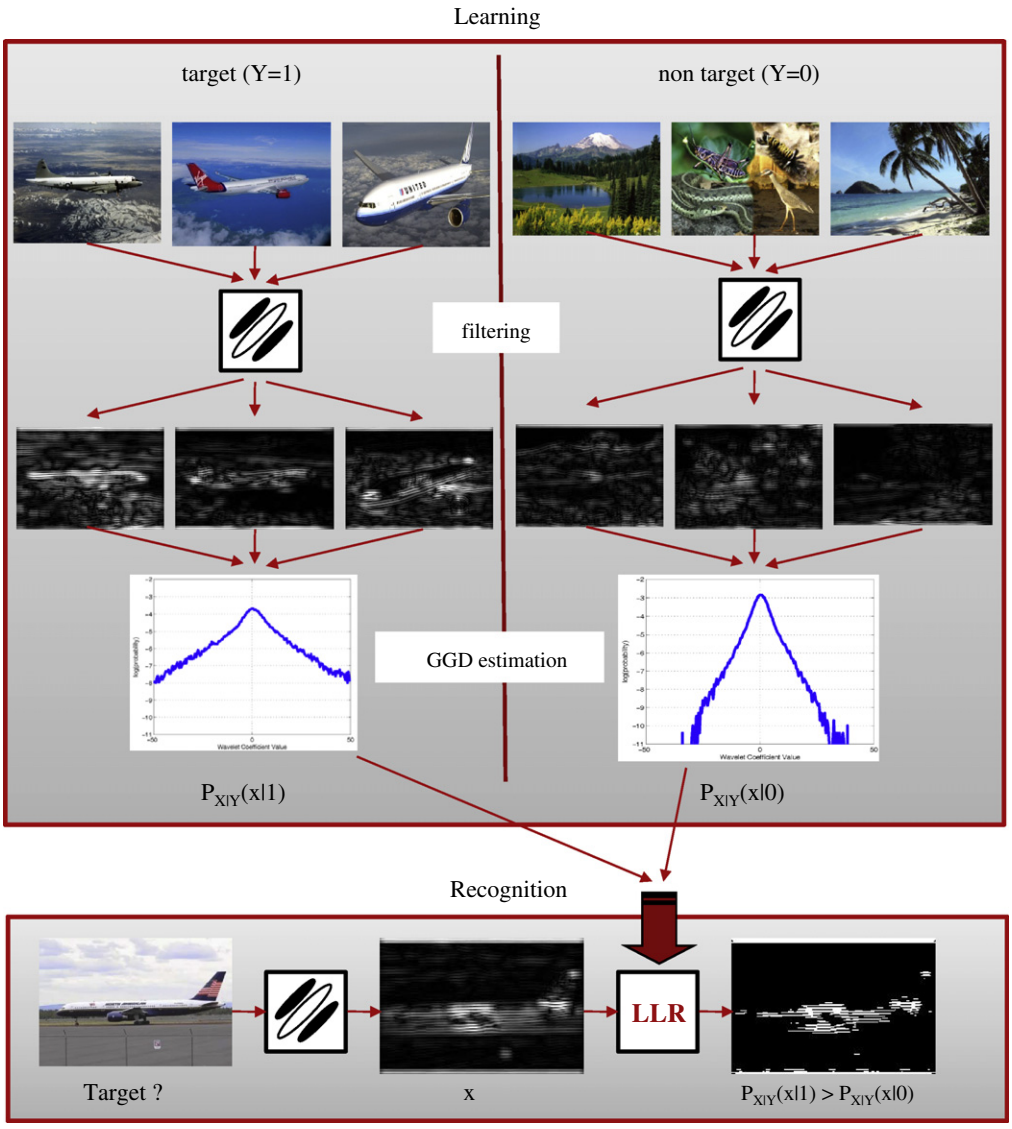

Fig. 2. Object recognition with the LLR measure. The learning stage is shown at the top of the figure. Gabor filtering is applied to examples of the target and null class. In this example, the target is the class of airplane objects. The probability distributions of the filter responses are then modelled with the GGD distribution. This enables the detection of objects from the target class in previously unseen images, as shown at the bottom. Given the filter responses to an unseen image, and the GGD estimates learned during training, the LLR is computed at each location of the visual field. Simple thresholding of this measure produces a binary map indicating the region of the vision field covered by the object.

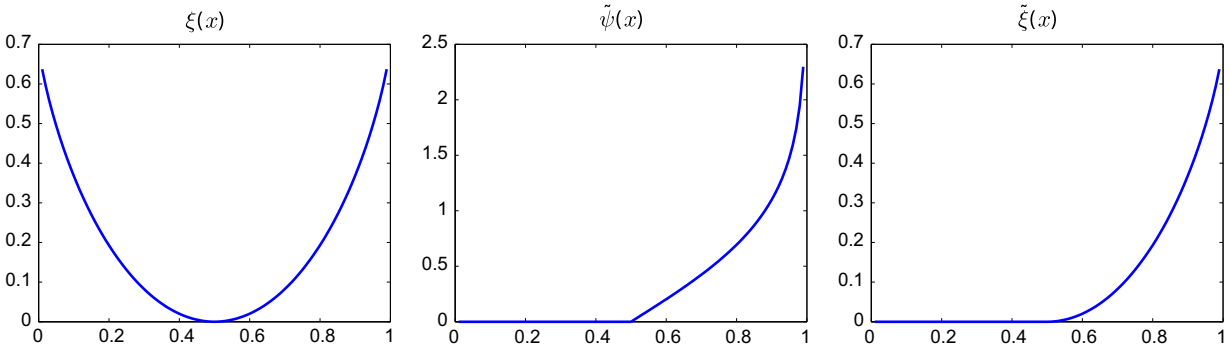

Fig. 3. Non-linear transformations of the posterior target probability that produce the information $I(Y; X = x)$ (left) and the confidence measures LLRC(x) (center) and IC(x) (right).

The third section of Table 1 addresses the characterization of the random variable X. This enables tasks like (1) feature selection, e.g. the identification of the most discriminant Gabor filters for a particular detection problem or (2) the determination of the entro-

py of X, e.g. to evaluate the uncertainty of the feature responses. This characterization usually requires the computation of empirical averages of the statistical inference operators discussed above, from a sample of observations $\mathscr{R} = \{x_1, \ldots, x_n\}$. Such averages are empirical estimates of popular statistical risks, which are referenced in the right-most column. These include the entropy

$$H[X] = -\int P_X(x)\log P_X(x)\,dx = E_X[-\log P_X(x)] \tag{9}$$

the mutual information

$$I(Y;X) = \sum_i \int P_{X,Y}(x,i)\log\frac{P_{X,Y}(x,i)}{P_X(x)P_Y(i)}\,dx = E_X[I(Y;X=x)] \tag{10}$$

or the Kullback-Leibler (KL) divergence

$$\mathrm{KL}[P_X(x)\|Q_X(x)] = \int P_X(x)\log\frac{P_X(x)}{Q_X(x)}\,dx = E_X\left[\log\frac{P_X(x)}{Q_X(x)}\right] \tag{11}$$

Once again, each inference principle leads to a different risk. For example, the expected LLR is a difference of two KL divergences

$$\mathrm{KL}[P_X(x)\|P_{X|Y}(x|0)] - \mathrm{KL}[P_X(x)\|P_{X|Y}(x|1)] \tag{12}$$

while the expected value of the information measure $I(Y; X = x)$ is the mutual information $I(Y; X)$ between the observation X and the class label Y. Finally, it is also possible to rely on expectations of the confidence measures of (5) and (6). These can be seen as one-sided versions of the KL difference and mutual information, which only average sample points identified as belonging to the target class (by the Bayes decision rule). Such averaging is equivalent to computing expectations with respect to the target class conditional distribution $P_{Y|X}(x|1)$, rather than $P_X(x)$. It, for example, simplifies the KL difference of (12) into the more standard KL divergence $\mathrm{KL}[P_{X|Y}(x|1)\|P_{X|Y}(x|0)]$. Again, their one-sided nature makes these risks particularly effective for top-down problems, such as target detection or recognition.

All risks based on KL divergences or mutual informations measure the discriminant power of X for target detection, and can be used for feature selection. When $\mathbf{X} = \{X_1,\ldots,X_k\}$ is a set of bandpass features, the dependencies of the feature responses to natural images tend to carry little information about the class label (Vasconcelos & Vasconcelos, 2009). This can be exploited to simplify the joint mutual information of the features with the class label into

$$I(\mathbf{X};Y) \approx \sum_k I(X_k;Y) \tag{13}$$

and justifies the computation of the overall discriminant power of $\mathbf{X}$ by adding the discrimination measures derived from each feature channel. We use this procedure to integrate the empirical risks of Table 1 across feature channels.

2.3. Inference under the GGD

When X follows a GGD, the computations above can be simplified into the form shown in the third column of Table 1. Here, all equations assume that X is either a GGD random variable of parameters (α, β), or a GGD random variable when conditioned on the class Y. In this case, the class conditionals $P_{X|Y}(x|i)$ have parameters (α_i, β_i), $i \in \{0, 1\}$. It is also assumed that $P_Y(0) = P_Y(1) = 1/2$, but this could be generalized into any label distribution. As noted by Gao and Vasconcelos (2009), the form of the negative log-likelihood

$$l_\alpha^\beta(x) = \frac{|x|^\beta}{\alpha^\beta} + K \tag{14}$$

is a straightforward consequence of (1). It follows that large values of $|x|$ indicate the locations of visual stimuli of low probability within the field of view. This is illustrated in Figs. 4 and 5a–c, which present two images, the magnitude $|x|$ of their convolution with a Gabor filter, and the NLL $l_\alpha^\beta(x)$ for the MAP GGD fit with $\beta = .5$. Note that the latter emphasizes details of the object or background which have very distinctive appearance from the rest of the image. In this sense, the log-likelihood operator behaves as an interest point operator, similar to a number of interest point operators currently popular in computer vision (& Stephens, 1988; Kadir & Brady, 2001; Mikolajczyk & Schmid, 2004; Sebe & Lew, 2003).

By definition, the LLR is a difference of two negative log likelihoods. It can be written as

$$g(x) = \log\frac{P_{X|Y}(x|1)}{P_{X|Y}(x|0)} = \left(\frac{|x|}{\alpha_0}\right)^\beta - \left(\frac{|x|}{\alpha_1}\right)^\beta + T, \tag{15}$$

where $T = \log\left(\frac{\alpha_0}{\alpha_1}\right)$. Figs. 4d and 5d show the LLR for motorbike detection on the images of (a). In both cases, α_1 was learned from a collection of bike images (target hypothesis), and α_0 from a random collection of natural images (null hypothesis). The LLR emphasizes the region of the motorbike, which is approximately uniformly highlighted, and inhibits the background.

Simple application of Bayes rule leads to the well known relation

$$P_{Y|X}(1|x) = \frac{1}{1+\frac{P_{X|Y}(x|0)}{P_{X|Y}(x|1)}} = \sigma[g(x)]$$

where $\sigma(x) = (1 + e^{-x})^{-1}$ is the sigmoid function. Hence, the target posterior is a sigmoidal transformation of the LLR. Similarly, (4) can be written as $I(Y;X = x) = \xi[P_{Y|X}(1|x)]$, with the non-linearity

$$\xi(x) = \log 2 + x\log x + (1-x)\log(1-x)$$

shown in Fig. 3. The application of these non-linearities to the images of Figs. 4d and 5d are shown in Figs. 4 and 5(e–f). They remap the LLR into the range [0–1]. While Gao and Vasconcelos (2009) have combined $\sigma(x)$ and $\xi(x)$ into a single non-linearity, there are non-trivial benefits in decoupling the two components. Note, in particular, that while the sigmoidal transformation maintains the emphasis on the bike region, the non-linearity associated with the information measure re-emphasizes some of the background. This is due to the fact that the latter is insensitive to the sign of the LLR (or, equivalently, to the sign of $P_{Y|X}(1|x) - 1/2$). In a strict information theoretic sense, the absence of an object is as *informative* as its presence for object detection (the classifier is simply very confident in the assignment of the image pixels to the background class). This is, however, undesirable for object detection, where the role is to detect object, and not background. When the two non-linearities are decoupled, this problem can be corrected by resorting to the measures of classification confidence of (5) and (6), which can be computed by composition of the sigmoid with the non-linearities of (7) and (8). The result, shown in Figs. 4 and 5(g–h) is a strong suppression of regions that belong to the background. This suppression enables very non-trivial gains in recognition accuracy, as will be shown in Section 3.3. Finally, all empirical risks can be computed by averaging some combination of these non-linearities. In summary, as noted in Table 1, most operations of statistical inference with GGD stimuli are non-linear mappings of the LLR $g(x)$ of (15).

2.4. Biological plausibility

Gao and Vasconcelos (2009) have shown that, given a sample $\mathscr{R}$ from a GGD distribution and using the estimate of (2) in (14),

$$l_\alpha^\beta(x) = \kappa\frac{|x|^\beta}{\sum_j |x_j|^\beta + v} + K \tag{16}$$

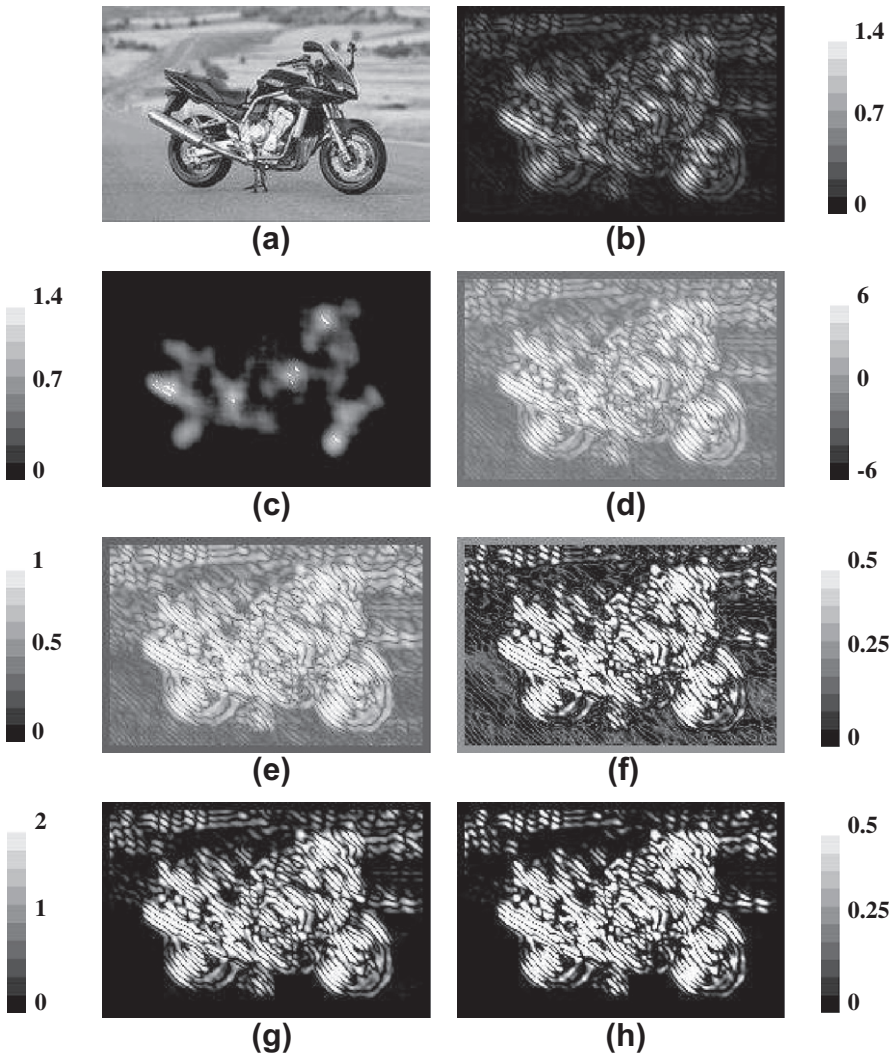

Fig. 4. (a) An image, (b) magnitude of Gabor responses, (c) NLL, (d) LLR, (e) target posterior probability, (f) information $I(Y; X = x)$, (g) LLRC(x), and (h) IC(x). The bars on the side of each image show the range of values corresponding to the pixel amplitudes.

The absolute value of x can be computed by half-wave rectification, i.e. as $|x| = x^+ + x^-$ where $x^+ = \max(x, 0)$ and $x^- = \max(-x, 0)$. This leads to the sequence of computations attributed to simple cells by the standard neurophysiological model of V1 (Carandini et al., 1997, 2005; Heeger, 1992; Hubel & Wiesel, 1962): linear filtering to produce a filter response x, half-wave rectification, and divisive normalization by the responses of other cells. For simplicity, we omit the decomposition into the rectified components (x^+, x^-) from all equations and network diagrams, working with $|x|$ instead. The combination of absolute value and divisive normalization as in (16) has recently been found to substantially improve the recognition accuracy of classical convolutional networks (Jarrett, Kavukcuoglu, Ranzato, & LeCun, 2009; Pinto, Cox, & DiCarlo, 2008; Pinto, Doukhan, DiCarlo, & Cox, 2009). However, no principled justification has been given for the importance of these operations. The discussion above suggests that this importance follows from their interpretation as estimators of the fundamental quantity of statistical inference (log–probability). The network representation of the simple cell is shown in Fig. 6.

Since the LLR is the difference of two log-probabilities, given two samples $\mathscr{R}_0$ and $\mathscr{R}_1$ from the null and target class, respectively, it follows that

$$g(x) = \frac{|x|^\beta}{\frac{1}{k}\sum_{x_j \in \mathscr{R}_0} |x_j|^\beta + \nu} - \frac{|x|^\beta}{\frac{1}{k}\sum_{x_j \in \mathscr{R}_1} |x_j|^\beta + \nu} + T \tag{17}$$

This leads to the biologically plausible implementation of the LLR with the network of Fig. 7. The main difference with respect to the network of Fig. 6 is that the filter responses are now differentially normalized by the units in the two dashed boxes. These boxes pool the response of other cells in a region $\mathscr{T}$ where the training sample $\mathscr{R}$ is collected. The bottom (top) units collect positive (negative) examples, producing an estimate of the GGD scale for the target class (null hypothesis). The region $\mathscr{T}$ localizes the cell computations. If $\mathscr{T}$ is the entire field of view, the GGD models are average distributions for the feature responses across the latter. For smaller $\mathscr{T}$, the cell response is tuned to the statistics of a sub-region of the field of view. Hence, the LLR can be computed by a differentially normalized simple cell. This prompted (Gao & Vasconcelos, 2009) to propose the LLR network as a model for simple cells. There are, however, two significant advantages in further including a sigmoidal non-linearity at the network output, as is now proposed in Fig. 7. First, this turns the cell into an estimator of the posterior target probability $P_{Y|X}(1|x)$, a more central quantity to the computations of Bayesian

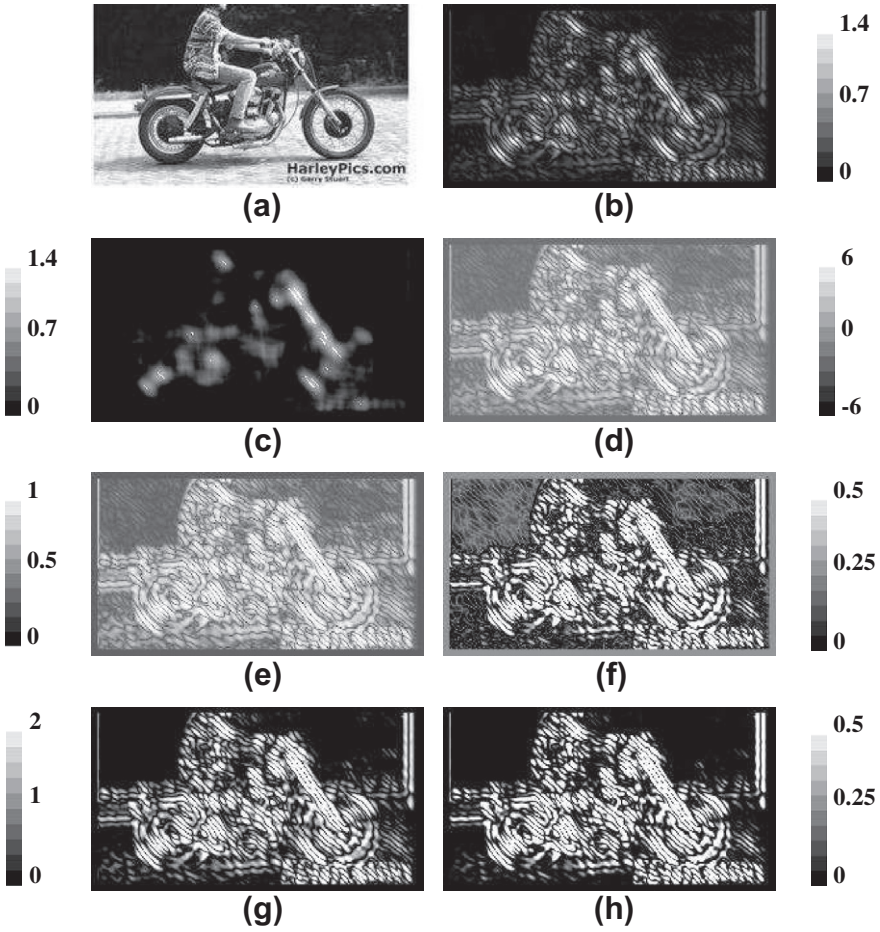

Fig. 5. (a) An image, (b) magnitude of Gabor responses, (c) NLL, (d) LLR, (e) target posterior probability, (f) information $I(Y; X = x)$, (g) LLRC(x), and (h) IC(x).

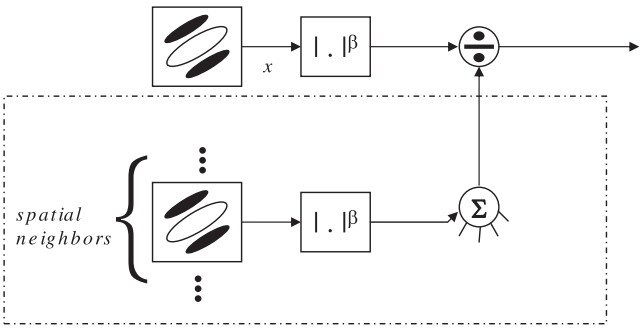

Fig. 6. The NLL is computed by a simple cell that normalizes a feature response x by the responses of its spatially neighboring units.

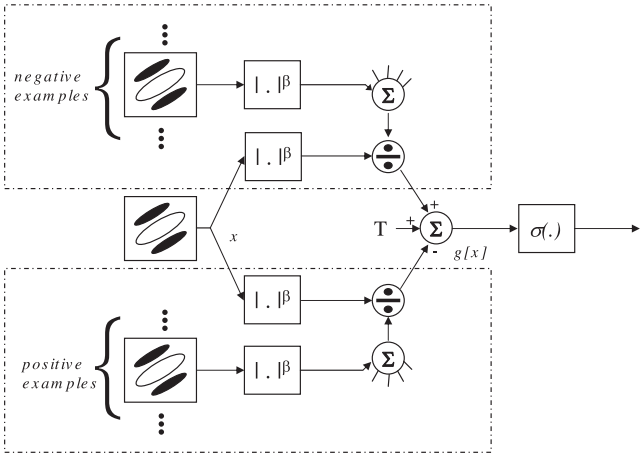

Fig. 7. A LLR unit divisively normalizes a feature response x differentially, using the outputs of two units that estimate GGD parameters under the target and null hypothesis. With the inclusion of the output non-linearity $\sigma(\cdot)$, this unit computes posterior target probabilities.

decision theory than the LLR. Second, it strengthens the biological plausibility of the simple cell model, by accounting for the saturation effects that are well known to hold for simple cell outputs, but are not replicated by the LLR.

Most risk estimates in the lower third of Table 1 consist of pooling some non-linear transformation of the posteriors $\sigma[g(x_i)]$, within some region $\mathscr{R}$ of the field of view. This makes the associated computations good candidates for complex cells. An example is the MI, for which the pooling operation is represented in Fig. 8. This network pools the responses of its afferent simple cells, after passing them through the non-linearity $\xi(\cdot)$. As shown in Fig. 3, this non-linearity is very close to quadratic, making the network a very good approximation of the standard energy model of complex cells by Adelson and Bergen (1985). The remaining empirical expectations of Table 1 can be implemented by replacing $\xi(\cdot)$ with the non-linearities $\tilde{\psi}(\cdot)$ or $\tilde{\xi}(\cdot)$, also shown in Fig. 3. The only exception

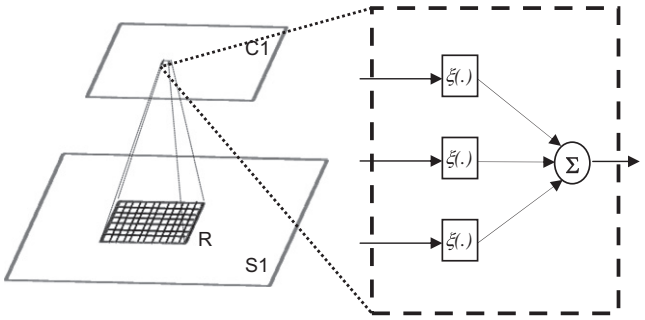

Fig. 8. A complex unit pools the responses of simple units within some region $\mathscr{R}$, after passing them through a non-linearity.

is the entropy network, which does not rely on the LLR $g(x)$. In this case, the complex cell pools the response of the NLL units in $\mathscr{R}$.

2.5. Saliency

A number of proposals for the measurement of visual saliency can be implemented by the networks of Table 1. We consider two bottom-up saliency methods, based on the detection of rare features, and a top-down approach, discriminant saliency, which accounts for the classes of the objects to detect.

2.5.1. Detection of rare features

A number of authors have advocated the detection of features of low probability as a criterion for visual saliency (Bruce & Tsotsos, 2006; Rosenholtz, 1999; Zhang et al., 2008). As discussed above, this criterion can be implemented with the NLL unit of Fig. 6. The detection of low probability features is also closely related to the most popular strategy for the detection of *interest points* in computer vision. A number of detectors from this literature identify image structure such as corners (Harris & Stephens, 1988), locations of strong image derivatives (Mikolajczyk & Schmid, 2004), wavelet coefficients of large magnitude (Sebe & Lew, 2003), or local maxima of image entropy (Kadir & Brady, 2001) that have low probability of occurrence. The features that elicit a strong response by NLL units generalize all these types of structure. For example (see Table 1), the combination of NLL units with a complex cell that pools its afferents linearly measures the entropy of the underlying feature responses.

It should be noted, however, that NLL units are technically not feature detectors, since they only compute the likelihood of feature responses. One possibility to transform them into detectors is to consider a discriminant version, that tests two hypotheses. Under the null hypothesis, x follows a GGD distribution $P_X(x)$ of parameters (α, β) estimated from the visual field. Under the alternative hypothesis, x follows a non-informative distribution $P_X(x) \propto 1$. The likelihood ratio is $g(x) \propto -\log P_X(x)$ and the posterior $P_{Y|X}(1|x) = \sigma(-\log P_X(x)) = \sigma(|x|/\alpha + K)$. The null hypothesis is rejected when $\frac{|x|}{\alpha}$ is large, i.e. large responses are better explained by the non-informative distribution. This implies that such responses are rare within the field of view. From an implementation point of view, the discriminant unit is identical to the NLL of Fig. 6, with the addition of an output sigmoid. We denote this combination as a *rare feature detector* (RFD).

2.5.2. Discriminant saliency

Discriminant saliency is defined with respect to a target and a *null* hypothesis. In the object detection context, the target is the class of objects to detect while the null hypotheses encompasses all stimuli outside that class. Locations of the visual field that can be assigned to the target class with minimal probability of error are declared salient, with degree of saliency equal to the classification confidence (Gao & Vasconcelos, 2009; Mahadevan, & Vasconcelos, 2007; Gao, Han, & Vasconcelos, 2009)

$$S(x) = \begin{cases} I(Y;X=x) & \text{if } P_{Y|X}(1|x) > .5 \\ 0, & \text{otherwise,} \end{cases} \tag{18}$$

This is the IC(x) measure of Table 1. If multiple responses $\{x_1,\ldots,x_K\}$ from feature X are available, the saliency of X is defined as $I(X;Y) = \frac{1}{K}\sum_i S(x_i)$, i.e. the expected confidence (MI) measure of the table. Saliency measurements derived from multiple feature channels are combined with (13). The last third of Table 1 suggests a number of other discriminant possibilities for measuring feature saliency: KL difference, mutual information $I(X;Y)$, or KL divergence. These measures differ from the expected confidence (MI), adopted by discriminant saliency, in relatively small details (mostly non-linearities). Such details could nevertheless be of consequence. For example, Jarrett et al. (2009) has found that simply taking the absolute value of the output of each unit of a classical convolutional network can produce drastic improvements in its recognition accuracy. The discussion above shows that these details can also completely alter the semantics of the network computations. For example, unlike the expected confidence (MI), the MI does not emphasize feature presence and could identify as salient a feature that is always absent from the target class. This is desirable for bottom-up saliency (Gao & Vasconcelos, 2009) but not necessarily for top-down applications, such as object detection or localization. We evaluate the performance of these measures in the following section, where it is shown that the choice of non-linearities can indeed have a significant impact on recognition performance.

3. Results

HMAX networks emulate the organization of the visual system by a cascade of two layers of simple and complex cells. We investigated the role of saliency in recognition by replacing the first HMAX layer with a saliency network. Under HMAX, this layer is quite simple: simple units perform filtering, and complex units pool simple unit responses within a spatial neighborhood, using a maximum operator. While these simple units have no probabilistic interpretation, max-based complex units are an interesting alternative to the sample averages of Table 1. They act more like a feature selection mechanism: rather than averaging responses, max-based pooling identifies the location of most salient response. This appears natural for detection-based saliency measures, e.g. the RFD. By replacing the first HMAX layer with a saliency network we can thus investigate three questions:

1. Is saliency important for visual recognition?
2. How do the various saliency criteria compare on an objective task, such as object recognition?
3. Is there an advantage in using max vs. the classical linear pooling?

In the broader neural network literature, there have been recent showings that some details of the network computations, e.g. what type of non-linearities or normalization is performed, can have a substantial impact in recognition accuracy (Jarrett et al., 2009; Pinto et al., 2009). As discussed above, the statistical interpretation of these operations makes it possible to assign semantics to all computations, with respect to optimality principles for discrimination, statistical inference, measurement of information, etc. This enables a more efficient search for optimal computations than trial-and-error (Jarrett et al., 2009), or brute-force optimization (Pinto et al.,

2009). To study these questions we performed a number of experiments, which are discussed in the remainder of this section.

3.1. Experiments

We start with a simple synthetic problem that provides intuition on the benefits of top-down discriminant saliency for recognition, and then present more extensive experiments on the Caltech101 benchmark, commonly used to evaluate object recognition performance. All experiments were based on the HMAX network, whose first layer was replaced by a saliency network. On Caltech101 we tested all saliency measures in the lower third of Table 1, as well as RFD, and the saliency detector of Itti, Koch, and Niebur (1998). For completeness, we also evaluated the use of a classical sigmoidal layer (no complex units or pooling, simple units a combination of filtering and a sigmoid) in the first HMAX layer, and the HMAX network itself. To investigate the advantages of max over linear pooling, all saliency networks were implemented with both. On the synthetic experiment we compared an HMAX network, HMAX with first layer replaced by a bottom-up saliency network of RFD units (HMAX + RFD), and HMAX with first layer replaced by a top-down saliency network of expected confidence(LLR) units (HMAX + EC).

In all experiments, for saliency units that involve divisive normalization, the pooling region $\mathscr{T}$ of the normalizing units was the whole image. In the case of bottom-up saliency (NLL or RFD units) the normalization is performed on-line, i.e. dividing by neighboring responses to the image to recognize. For top-down saliency (LLR units) the normalizing coefficients are learned during training, when the network is exposed to images from the target and null hypotheses. For complex units, the pooling region $\mathscr{R}$ was as specified in Mutch and Lowe (2008).

The second layer of the HMAX network consists of a set of radial basis function (RBF) units, centered at prototypes randomly sampled from the responses of the first HMAX layer, during training. On Caltech101 we used the implementation of Serre, Wolf, Bileschi, Riesenhuber, and Poggio (2007), which includes 4075 RBF units. On the synthetic experiment we used a smaller network of 100 units. For LLR units, training produces two divisive normalization parameters (α_i^β) per object class. For a given RBF prototype **P**, the parameters of the afferent simple units are set to the values $\alpha_i^\beta(\mathbf{P})$ with which **P** was learned (i.e. the parameters learned from the image class which originated **P**). Other than these modifications, the network is exactly as described in Mutch and Lowe (2008).

3.2. Synthetic problem

To gain some insight on the role of discriminant saliency in recognition, we considered the simple problem of learning to differentiate underlined from non-underlined characters. This was formulated as a two-class recognition problem, involving the stimuli of Fig. 9. Each network was trained with the top two images of the figure, using underlined Xs as examples from the target class, and regular Xs as example non-targets. This made the classes identical up to a salient feature of the underlining concept (the underline bar). The network was then used to classify 20 test images, containing either targets or non-targets. To increase the difficulty of the task, the character used on the test images (Y) was different from that used for training (X), and random noise was added to all images.

The recognition accuracy achieved by the three networks was 90% for HMAX + EC, 55% for HMAX+RFD, and 50% for HMAX. The superior performance of the network with top-down saliency can be understood by analyzing the intermediate network responses, shown in Figs. 9 and 10. Consider the response of the first network layer, shown in Fig. 9. The HMAX network only has access to Gabor filter responses, which are very similar for target and non-target. This makes it very difficult for the subsequent HMAX stages to distinguish between the two classes. Because none of the parts of the underlined Xs pop-out within the target displays, the saliency response of RFD is basically a contrast enhanced version of the filter responses. This does not improve the recognition accuracy substantially, since contrast variability is not the reason for the poor performance of HMAX on this classification problem (although it can be a source of concern for problems involving natural images where, as we will see in the next section, HMAX + RFD tends to

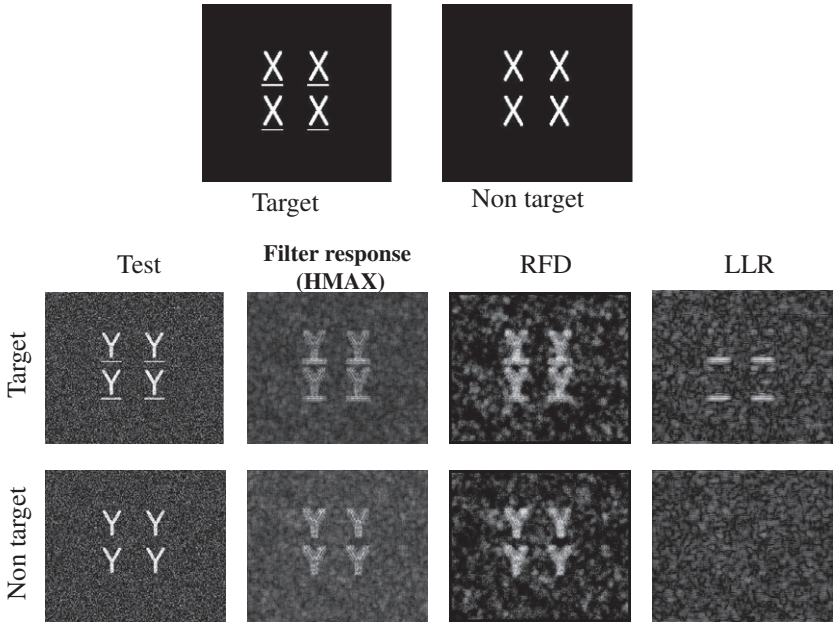

Fig. 9. Detection of underlined characters. *Top row*: Training examples from target and non-target class. *Bottom rows*: Examples of test stimuli from the target and non-target class, and layer 1 responses from the three networks considered.

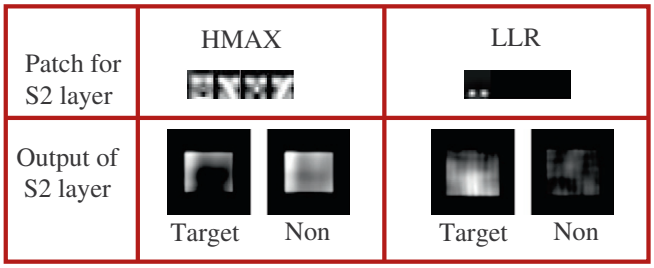

Fig. 10. *Top*: Most discriminant filter (the four orientation channels are shown) of the second network layer, for HMAX (left) and HMAX + EC (right). By most discriminant it is meant that this is the filter given larger weight by the linear SVM classifier at the network output. *Bottom*: Example output of the simple cells in layer 2, to target and non-target stimuli.

outperform HMAX). Hence, the performance of HMAX and HMAX + RFD is basically identical.

The underline bar is, however, salient in the top-down sense, since it is the only part that distinguishes the target and non-target examples. Because the units of the HMAX + EC network compute the LLR between target and non-target hypothesis, they produce a strong response to underline bars (plausible under target, but not plausible under the non-target hypothesis) and a weak response to everything else (equally plausible, or non-plausible, under the two hypotheses). The network has thus learned that horizontal bars are discriminant features for the detection of underlined characters, and thus salient. Its first layer acts as a detector of these bars, and its very different responses to targets and non-targets are easily detected by the subsequent network stages. Fig. 10 presents the most discriminant filter of the second layer (four orientation channels shown), for the HMAX and HMAX + EC networks. Note how the filter of HMAX + EC is a detector of horizontal bars, a property that does not hold for the other networks. In result, the output of the second layer of HMAX + EC is uniformly large for underlined carachters, and almost null for non-targets. This is unlike the other two networks, whose second layers respond to both targets and non-targets. It is thus not surprising that HMAX + EC achieves a substantially higher recognition accuracy.

3.3. Caltech101 experiments

To evaluate the impact of the various saliency principles on the classification of natural images, we performed a number of experiments on Caltech101. All experiments were based on the experimental protocol of Mutch and Lowe (2008). We considered the multiclass recognition task, where 30 images per class are used for training and a maximum 50 of the remaining for test. In all experiments the reported recognition rate is the average over five independent runs, with different train and test sets (randomly sampled images). Table 2 presents the recognition accuracy achieved with each variant of the first network layer. A graphic display of these rates, as well as the associated error bars, is shown in Fig. 11.

A few interesting observations can be made. First, the two expected confidence criteria achieve the best results. Their performance is similar, but EC(LLR) attains slightly higher recognition rates. These methods can be implemented with simple units that compute the target posterior probability, i.e. a combination of a differentially and divisively normalized (LLR) unit and a sigmoid $\sigma(\cdot)$. The gains with respect to the remaining networks can be very significant. Second, saliency criteria based on rare features (ENLL and RFD) perform worse than saliency criteria based on discrimination (the expected confidence measures). On the other hand, both rare feature criteria have clearly better performance than sigmoid or HMAX. This suggests that rare feature (interest point) detection can be useful when statistics of the target object class are not available. Note that, under the rare feature criteria, none of the two network layers requires class-specific training. While the same holds for the saliency detector of Itti et al. (1998), its performance (51.8%) is substantially weaker than those of ENLL or RFD.

Third, the "one-sided" confidence measures EC(LLR) and EC(MI) perform substantially better than their "two sided" counterparts, such as the ELLR or the MI used in Gao and Vasconcelos (2009). This implies that the choice of non-linearities (e.g. $\tilde{\xi}$ instead of ξ or $\tilde{\psi}$ instead of ψ) can have a very non-trivial impact in recognition accuracy. It appears to be particularly important for the cells to fire only when the target is present. Fourth, for most networks, max-based pooling has inferior performance to averaging. This implies that it is important to fully characterize features, and not only select locations where they are informative for the classification. The only

Table 2
Recognition rates on Caltech101, using 30 training examples per class. All abbreviations are the same as in Table 1. Furthermore, EC means expected confidence, ELLR expected LLR, ENLL expected NLL, RFD rare feature detection.

Network	Simple units			Complex units						
	Divisive normalization			Non-linearity				Pooling		
	NLL	LLR	$\sigma(\cdot)$	$\xi(\cdot)$	$\psi(\cdot)$	$\tilde{\xi}(\cdot)$	$\tilde{\psi}(\cdot)$	Sum	Max	Accuracy
EC (LLR)		√	√				√	√		61.9
		√	√				√		√	58.2
EC (MI)		√	√			√		√		60.3
		√	√			√			√	53.1
ELLR		√	√		√			√		54.6
		√	√		√				√	53.1
MI		√	√	√				√		50.3
		√	√	√					√	52.3
ENLL	√							√		58.2
	√								√	56.8
RFD	√		√					√		55.1
	√		√						√	55.2
Itti et al. (1998)	–	–	–	–	–	–	–	–	–	51.8
Sigmoid			√							42
HMAX								√		40.5
									√	43.4
			√					√		44.1
			√						√	46.6

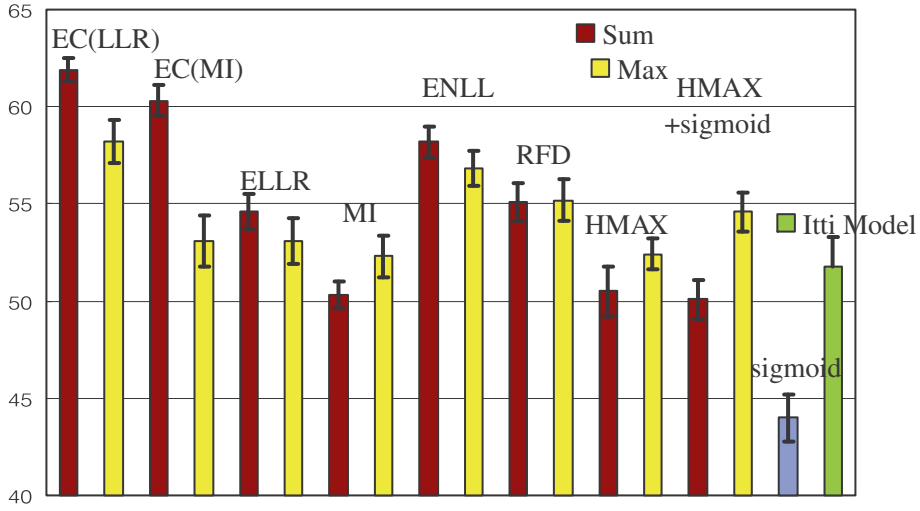

Fig. 11. Recognition rates on Caltech101, using 30 training examples per class.

network for which max pooling consistently achieves better performance is HMAX (where the lack of sophistication of the simple units makes the network with average pooling linear). Furthermore, max-based pooling is prone to large performance variability. For example, the EC(MI) network drops from 60% to 53% recognition rate when averaging is replaced by max pooling. Finally, the classical sigmoid layer has the worst performance of all considered. However, the simple addition of a pooling stage can improve performance considerably, especially when combined with max pooling.

3.4. Comparison to state-of-the-art results

To the best of our knowledge, the current state-of-the art results for object recognition with HMAX networks are those presented in Mutch and Lowe (2008). This work reported significant improvements over the base HMAX performance, through a number of enhancements to the original network. Some of these involved additional training, e.g. to select features, others are heuristics that were shown to improve performance. Table 3 presents the contributions by these enhancements, as reported in Mutch and Lowe (2008). As can be seen from the table, the simple use of the saliency layer, *without any further optimization,* outperforms the gains of *all* enhancements of Mutch and Lowe (2008). One of these improvements is a feature selection stage. Rather than using 4075 randomly sampled prototypes, a larger set of 12,000 are collected. The network is trained with this larger set, and a support vector machine is used to select the most discriminant 4075. When we retrained the network containing the saliency layer in this manner, the performance increased to 64%, as opposed to the 56% reported by Mutch and Lowe. While we have not yet experimented with any of their other suggestions, or performed any other optimization, these results suggest that the inclusion of saliency can significantly boost the performance of feedforward object recognition.

Table 3
Multiclass classification results for 101 categories.

Model	15/cat	30/cat
Base model of Serre et al. (2007)	33	42
+ sparse S2 inputs Mutch and Lowe (2008)	35	45
+ inhibited S1/C1 outputs Mutch and Lowe (2008)	40	49
+ limited C2 invariance Mutch and Lowe (2008)	48	54
+ feature selection Mutch and Lowe (2008)	51	56
EC (LLR) with sum pooling	56	62
+ feature selection described in Mutch and Lowe (2008)	58	64
Convolutional net of Pinto et al. (2008)	–	42
+ second HMAX layer	–	56
Convolutional net of Jarrett et al. (2009)	–	56
+ random filters	–	63
+ unsupervised filters	–	64
+ back-propagation filters	–	66
Lazebnik et al. (2006)	56	65
Zhang et al. (2006)	59	66

In the broader area of convolutional networks, recent studies have addressed the role of non-linearities and normalization in object recognition (Jarrett et al., 2009; Pinto et al., 2008). These works advocate the use of divisive normalization as a form of contrast normalization, that improves the robustness of the neural network when trained from small samples, as is the case of Caltech101 (Jarrett et al., 2009). This is a strictly bottom-up explanation for the role of divisive normalization, and comparable to the ENLL and RFD saliency measures discussed in this work. Comparison with these methods should be performed with care, since the network parameters are not the same. For example, while it has become somewhat popular to claim that method of Pinto et al. (2008) beats the state-of-the-art in computer vision, the truth is that its implementation is far from the standard in this area. For example, while (for computational efficiency) most computer vision implementations rely on a relatively small set of filters (e.g. Gabor filters at four orientations) and a relatively small number of network outputs (4075 for the first HMAX network (Serre et al., 2007), 12,000 for enhanced HMAX (Mutch & Lowe, 2008)), this method relies on a much larger filter set (12 orientations), and a much larger output dimensionality (86,400–116,400). The network has a single layer and is complemented by a classifier that combines a principal components analysis of very disputable biological plausibility, and an SVM. While the recognition accuracy originally reported by the authors is of 65% (30 images per category), our implementation with (1) the Gabor filter front-end and (2) the output dimensionality used by the HMAX networks only achieved 42%. Further inclusion of the second HMAX layer raised recognition performance to 56%. We note that this is consistent with the results of Table 2, as the network of Pinto et al. (2008) is similar to the RFD network. Hence, it is not surprising that the results are in between those of ENLL (58.2%) and RFD (55.1%).

Similar performance was documented by Jarrett et al. (2009), who have obtained 55.8% accuracy with a two layer network

including divisive normalization in the two layers (as opposed the one we tested, where only the first layer was modified). This work has tested a number of extensions, including the use of filters learned from the training data, in both a bottom-up and top-down manner. All results reported are lower than those achieved with the EC(LLR) network, except when the filters are trained in a discriminant manner. Note that, in this case, the convolutional network has two layers of trained filters and divisive normalization, network training is orders of magnitude more complex than that required by the saliency network (back-propagation for the former vs. the individual tuning of the divisive normalization weights of each simple cell, according to (2), for the latter), and the gains are very marginal (65.5% vs. 64%). The filters of the EC(LLR) network could also have been trained in a discriminant manner, but we have not attempted to perform this optimization.

For completeness, we also report the state-of-the-art results on Caltech101 from the broader recognition literature in computer vision, where biological plausibility is not a constraint. We consider here only methods that use a single image representation, and are therefore comparable to the networks proposed above. In this class, the best performance in the literature is in the range of 65–66% (Lazebnik, Schmid, & Ponce, 2006; Zhang, Berg, Maire, & Malik, 2006) and barely superior to the 64% now reported for the biologically plausible networks. Obviously, better performance should be attainable by combining multiple image representations, e.g. by adding features that capture color or shape properties to the set of Gabor functions that we consider in this work. This is indeed a popular strategy in the computer vision literature, where it has been shown that substantial improvements over (Lazebnik et al., 2006; Zhang et al., 2006) can be achieved with support vector machines combining multiple kernels (Gehler & Nowozin, 2009; Varma & Ray, 2007). Such combinations of multiple image representations could also be applied to the networks that we have proposed, but are beyond the scope of this work.

4. Discussion and conclusion

Overall, the results presented above support three main conclusions:

- saliency (attention) has a significant positive impact on recognition,
- but this impact is largest when saliency is discriminant (of a top-down nature). Unsupervised learning of interest points does not perform as well, although it consistently achieves better performance than no saliency at all (standard HMAX);
- max-based pooling does not appear to have an advantage over averaging, indicating that selecting discriminant features is more important than locating them exactly.

It could be argued that replacing the raw filter outputs with discriminant saliency measures is simply a form of normalization, whose benefits have already been pointed out in the literature (Jarrett et al., 2009; Pinto et al., 2008). While normalization has advantages of its own, as shown by the gains of the ENLL and RFD networks over their sigmoidal counterpart, this is not the whole story. The results above show that non-trivial additional gains can be obtained with *intelligent normalization,* which tunes the cell responses according to the target recognition class, at a very marginal cost in computation. This is a top-down saliency operation.

To illustrate the benefits of this type of saliency for classification of natural images, we examined the intermediate computations of the different types of networks. Fig. 12 shows the output of the saliency layer for an example image of the "accordion" class. The figure shows the saliency maps produced for four Gabor orientation channels. The first row presents the magnitude $|x|$ of the Gabor responses (no saliency processing), the second row the output of the NLL units (bottom-up processing), and the third row that of the LLR units trained for accordion detection (top-down saliency). Note that both types of saliency units reinforce the contrast of certain areas of the image, leading to a more distinctive visual signature than the simple magnitude of Gabor responses. The responses of the two types of saliency units are, nevertheless, quite different. NLL has no knowledge of the accordion class, and simply highlights visual features that have low probability within the field of view. These tend to be the keyboards that appear on each side of the instrument. The diagonal edges, which are a distinctive pattern of the accordion object but plentiful on this image, are suppressed. This implies that there is some loss of information, a limitation of bottom-up saliency for recognition: universal saliency criteria

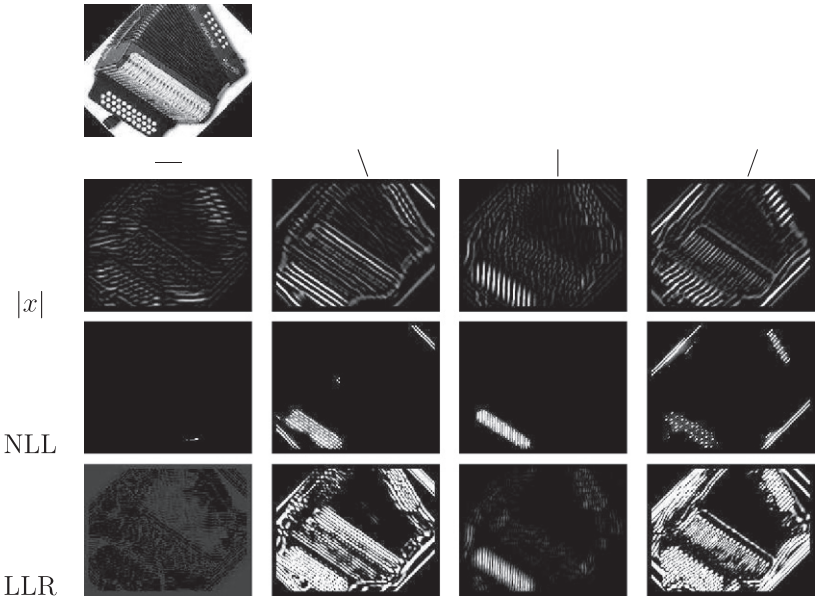

Fig. 12. An image from the "accordion" class, and corresponding saliency outputs for Gabor channels of four orientations. *Top row*: Magnitude of the Gabor responses. *Center*: Saliency maps produced by NLL units. *Bottom*: Saliency maps of LLR units.

(such as low probability, or contrast normalization) fail to capture the salient attributes that are *specific* to any given object class.

The top-down LLR units exhibit a substantially different behavior. For orientation channels that do not contain substantial discriminant information about the target class, they behave similarly to NLL units. However, for orientations that capture distinctive object patterns (such as the large density of parallel lines in the accordion class), they respond very strongly throughout the field of view, highlighting the whole object. The resulting saliency patterns are thus much more distinctive templates than those produced by Gabor filtering or NLL. When used in the second HMAX layer, these templates are much more discriminant for the target class, enabling better detection performance. In summary, the attributes that are salient for object recognition vary from one object class to another. The identification of such attributes requires top-down processing informed by the class structure associated with the recognition problem. Discriminant saliency implements this type of processing, leading to the extraction of intermediate features that are highly informative for object recognition. This results in higher recognition rates.

We finish by emphasizing one of the most interesting findings of this work: that subtle modifications to the computations of Gao and Vasconcelos (2009) can lead to substantial changes of network behavior. These include (1) obtaining good performance on top-down tasks such as recognition (rather than just bottom-up saliency), (2) computing new statistical quantities of interest, namely all measures of Table 1, (3) explaining properties such as simple cell saturation, and (4) assigning semantics to all network components. All of these help understand why network modifications that appear minor a-priori can have a dramatic impact in performance. For example, while Gao and Vasconcelos (2009) have shown that, among the non-linearities of Fig. 3, $\xi(x)$ performs best for bottom-up saliency, the results now presented show that $\tilde{\tilde{\xi}}(x)$ is clearly better for top-down saliency. This can be seen from Table 2, where replacing $\tilde{\tilde{\xi}}(x)$ by $\xi(x)$ leads to a *substantial* decrease of recognition accuracy, e.g. from 60.3% (EC(MI)) to 50.3% (MI).

Although the dramatic influence of non-linearities on recognition performance has been documented in the literature (Jarrett et al., 2009), it can be quite puzzling in the absence of the statistical interpretation now provided. Why would simply changing a non-linearity degrade the performance so much? And why does it matter so much that the non-linearity is "one sided"? The statistical interpretation clarifies this behavior: while the EC(MI) is a detector of target presence, the MI is equally happy to detect target presence or absence. The semantics of the network computations are, therefore, completely different. Under MI, the network produces large responses to background regions that can be classified as *either target or non-target* with high confidence. This increases the difficulty of target detection. On the other hand, under EC(MI) the network only produces large responses to regions that contain the target.

References

Adelson, E., & Bergen, J. (1985). Spatiotemporal energy models for the perception of motion. *Journal of the Optical Society of America, 2*(2), 284–299.

Barlow, H. (2001). Redundancy reduction revisited. *Network: Computation in Neural Systems, 12*.

Bruce, N., & Tsotsos, J. (2006). Saliency based on information maximization. *Neural Information Processing Systems*.

Buccigrossi, R., & Simoncelli, E. (1999). Image compression via joint statistical characterization in the wavelet domain. *IEEE Transactions on Image Processing*.

Carandini, M., Demb, J., Mante, V., Tolhurst, D., Dan, Y., Olshausen, B., et al. (2005). Do we know what the early visual system does? *Journal of Neuroscience, 25*.

Carandini, M., Heeger, D., & Movshon, A. (1997). Linearity and normalization in simple cells of the macaque primary visual cortex. *Journal of Neuroscience, 17*, 8621–8644.

Do, M., & Vetterli, M. m. (2002). Wavelet-based texture retrieval using generalized Gaussian density and Kullback–Leibler distance. *IEEE Transactions on Image Processing, 11*(2), 146–158.

Duda, R., Hart, P., & Stork, D. (2001). *Pattern classification*. John Wiley & Sons.

Friedman, J., Hastie, T., & Tibshirani, R. (2000). Additive logistic regression: A statistical view of boosting. *The Annals of Statistics, 28*.

Gao, D., Han, S., & Vasconcelos, N. (2009). Discriminant saliency, the detection of suspicious coincidences, and applications to visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Gao, D., & Vasconcelos, N. (2009). Decision-theoretic saliency:computational principles, biological plausibility, and implications for neurophysiology and psychophysics. *Neural Computation, 21*.

Gehler, P., & Nowozin, S. (2009). On feature combination for multiclass object classification. In *International conference on computer vision*.

Harris, C., & Stephens, M. (1988). A combined corner and edge detector. In *Alvey vision conference*.

Hastie, T., Tibshirani, R., & Friedman, J. (2001). *The elements of statistical learning; data mining, inference and prediction*. New York: Springer Verlag.

Heeger, D. (1992). Normalization of cell responses in cat striate cortex. *Visual Neuroscience, 9*.

Huang, J., & Mumford, D. (1999). Statistics of natural images and models. In *Computer vision and pattern recognition*.

Hubel, D., & Wiesel, T. (1962). Receptive field, binocular interaction, and functional architecture of in the Cat's visual cortex. *Journal of Physiology, 160*.

Itti, L., Koch, C., & Niebur, E. (1998). A model of saliency-based visual attention for rapid scene analysis. *IEEE Transaction on Pattern Analysis and Machine Intelligence, 20*.

Jarrett, K., Kavukcuoglu, K., Ranzato, M., & LeCun, Y. (2009). What is the best multi-stage architecture for object recognition? In *International conference on computer vision*.

Kadir, T., & Brady, M. (2001). Scale, saliency and image description. *International Journal of Computer Vision, 45*.

Lazebnik, S., Schmid, C., & Ponce, J. (2006). Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *Computer vision and pattern recognition*, June 2006.

Lecun, Y., Bottou, L., Bengio, Y., & Haffiner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*.

Linsker, R. (1988). Self-organization in a perceptual Network. *IEEE Computer, 21*(3), 105–117.

Mahadevan, V., & Vasconcelos, N. (2007). The discriminant center-surround hypothesis for bottom-up saliency. *Neural Information Processing Systems*.

Miau, K., Papageorgiou, C., & Itti, L. (2001). Neuromorphic algorithms for computer vision and attention. In *Proceedings of SPIE 46 annual international symposium on optical science and technology* (Vol. 4479, pp. 12–23).

Mikolajczyk, K., & Schmid, C. (2004). Scale and affine invariant interest point detectors. *International Journal of Computer Vision, 60*(1), 63–86.

Mutch, J., & Lowe, D. (2008). Object class recognition and localization using sparse features with limited receptive fields. *International Journal of Computer Vision, 80*, 45–57.

Pinto, N., Cox, D., & DiCarlo, J. (2008). Why is real-world visual object recognition hard? *PLoS Computational Biology, 4*(1).

Pinto, N., Doukhan, D., DiCarlo, J., & Cox, D. (2009). A high-throughput screening approach to discovering good forms of biologically-inspired visual representation. *PLoS Computational Biology, 5*(11).

Riesenhuber, M., & Poggio, T. (1999). Hierarchical models of object recognition in cortex. *Nature Neuroscience, 2*.

Rosenholtz, R. (1999). A simple saliency model predicts a number of motion popout phenomena. *Vision Research, 39*, 3157–3163.

Rumelhart, D., Smolenksy, P., Mcclelland, J., & Hinton, G. (1986). *Parallel distributed processing: Explorations in the microstructure of cognition*. MIT Press.

Sebe, N., & Lew, M. (2003). Comparing salient point detectors. *Pattern Recognition Letters, 24*(Jan), 89–96.

Serre, T., Wolf, L., Bileschi, S., Riesenhuber, M., & Poggio, T. (2007). Robust object recognition with cortex-like mechanisms. *IEEE Transaction on Pattern Analysis and Machine Intelligence*.

Srivastava, A., Lee, A., Simoncelli, E., & Zhu, S. (2003). On advances in statistical modeling of natural images. *Journal of Mathematical Imaging and Vision, 18*, 17–33.

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

Varma, M., & Ray, D. (2007). Learning discriminative power-invariance trade-off. In *International conference on computer vision*.

Vasconcelos, M., & Vasconcelos, N. (2009). Natural image statistics and low complexity feature selection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Walther, D., & Koch, C. (2006). Modeling attention to salient proto-objects. *Neural Networks, 19*, 1395–1407.

Yarbus, A. (1967). *Eye movements and vision*. New York: Plenum.

Zhang, H., Berg, A., Maire, M., & Malik, J. (2006). Svm-knn: Discriminative nearest neighbor classification for visual category recognition. In *Computer vision and pattern recognition*, June 2006.

Zhang, L., Tong, M., Marks, H., Tim, K., Shan, H., & Cottrell, G. (2008). SUN: A Bayesian framework for saliency using natural statistics. *Journal of Vision, 8*(7), 1–20.