Discriminant saliency, the detection of suspicious coincidences, and applications to visual recognition

Dashan Gao

Visualization and Computer Vision Lab

General Electric Global Research*,

Niskayuna, NY, 12309, USA

Sunhyoung Han Nuno Vasconcelos

Department of Electrical and Computer Engineering,

University of California, San Diego

La Jolla, CA 92093

* This work was performed while the author was with the Department of Electrical and Computer Engineering, UCSD

Abstract

A discriminant formulation of top-down visual saliency, intrinsically connected to the recognition problem, is proposed. The new formulation is shown to be closely related to a number of classical principles for the organization of perceptual systems, including infomax, inference by detection of suspicious coincidences, classification with minimal uncertainty, and classification with minimum probability of error. The implementation of these principles with computational parsimony, by exploitation of the statistics of natural images, is investigated. It is shown that Barlow's principle of inference by the detection of suspicious coincidences enables computationally efficient saliency measures which are nearly optimal for classification. This principle is adopted for the solution of the two fundamental problems in discriminant saliency, feature selection and saliency detection. The resulting saliency detector is shown to have a number of interesting properties, and act effectively as a focus of attention mechanism for the selection of interest points according to their relevance for visual recognition. Experimental evidence shows that the selected points have good performance with respect to 1) the ability to localize objects embedded in significant amounts of clutter, 2) the ability to capture information relevant for image classification, and 3) the richness of the set of visual attributes that can be considered salient.

Index Terms

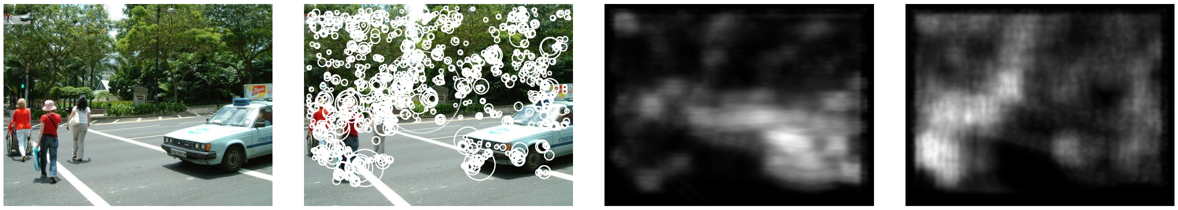
Visual saliency, interest point detection, coincidence detection, visual recognition, object detection from cluttered scenes, infomax feature selection, saliency measures, natural image statistics

I. Introduction

Biological vision systems have a remarkable ability to recognize objects under adverse conditions, such as highly cluttered scenes. The use of saliency mechanisms is believed to play an important role in this robustness to clutter. They make salient locations "pop-out", driving attention to the appropriate regions of the visual field [1], [2]. This enables organisms to focus their limited perceptual resources on the most pertinent subsets of the sensory stimuli, facilitating subsequent visual processing. In the biological world, vision systems rarely need to perform an exhaustive scan of a scene in order to detect an object of interest.

Saliency has been extensively studied in both the biological and computer vision literatures over the last decades. In biological vision, most research addresses the understanding of how attentional mechanisms work, either through psychophysics experiments in psychology, or neural recordings in neurophysiology. Although a tremendous amount of knowledge about saliency has been amassed in this way, this literature is not rich in computational models. When such models are proposed, they tend to focus on high-level justifications for specific attention mechanisms, and do not necessarily translate into computer vision algorithms. Although there are notable exceptions, such as the pioneering models of [3]–[6] which have been shown useful for computer vision [7]–[9], they frequently lack a formal justification, based on a unifying *computational principle* for saliency. In the absence of clearly defined optimality criteria, it is difficult to evaluate, in an objective sense, the goodness of the proposed algorithms, or to develop a computational theory, and algorithms, for optimal saliency.

In computer vision, most saliency research has focused on the extraction of image locations, called *interest points*, that exhibit some mathematically well defined properties, e.g. stability under geometrical transformations. These saliency detectors have been widely adopted in applications such as object tracking and recognition, and more recently, learning object detectors from weakly supervised (unsegmented) training examples [10]–[17]. In these applications, saliency is often justified as a pre-processing step that saves computation and improves robustness, facilitating the design of subsequent stages. Most saliency detectors in this category have clearly defined optimality criteria, which can be divided into two major classes. The first, and most popular, treats saliency detection as the optimal *detection of specific visual attributes*. Most detectors in this class have roots in the areas of structure-from-motion and tracking. The most prevalent examples are edges and corners [18], [19] but there have also been proposals for other low-level attributes, e.g. contours [20], [21], local symmetries [22], [23], and blobs [24]. These detectors can often be embedded in scale-space [25], to achieve invariance with respect to transformations such as scaling or affine mappings [26], [27]. The second class of detectors strives to be optimal with respect to more generic criteria, such as image *complexity*. For example, [28] defines saliency by the variance of Gabor filter responses over multiple orientations, while [29] equates saliency to the absolute value of the coefficients of a wavelet image decomposition,



(a) (b) (c) (d)

Fig. 1. Bottom-up vs. top-down saliency. (a) A scene with a car, people, trees and pavement. (b) salient points detected by a (bottom-up) Harris-Laplacian detector. (c,d) saliency maps generated by a top-down saliency detector trained to detect cars (c) and people (d). Bright pixels flag locations of high saliency.

and [30] to the entropy of the distribution of local intensities. These more generic principles are more flexible than those tied to specific visual features. They could declare as salient any of the low-level attributes discussed above, depending on the image under consideration.

Although both computational formulations have been widely adopted in object recognition, they do not tie the optimality of saliency judgements to the recognition goal. In result, the detected salient locations do not necessarily co-occur with the objects to be detected. This limitation is illustrated in Figure 1(a), for a scene containing people and a car, among other visual concepts (road pavement, trees, etc.). Figure 1(b) shows the salient locations, extracted from the image, by the Harris-Laplacian detector [26]. These locations are equally distributed over cars, people, and other concepts. For a recognizer of 'cars' or 'people', they are far from optimal, as the set of salient locations is far greater than the area occupied by the objects of interest. For a 'pavement' detector they are catastrophic: most regions of interest are not considered salient, and are discarded before detection ever takes place. In general, definitions of saliency divorced from the recognition problem cover the gamut between computational inefficiency and catastrophic failure. This has motivated various researchers to argue that saliency offers no advantage over either dense sampling, or random sub-sampling of image locations [13], [31], [32].

Such conclusions are, however, difficult to reconcile with the predominant roles of attention and saliency in biological vision. The most likely explanation is that biology resorts to two complementary saliency mechanisms: a *bottom-up* stimulus-driven component and a *top-down* stage driven by recognition goals [1], [2]. While the bottom-up component is emulated by the computer vision detectors, the top-down mechanisms are not. Top-down mechanisms can be seen as weak classifiers that drive attention to the regions of the visual field which are relevant for the recognition problem. This is illustrated in Figure 1 (c) and (d), which present the saliency maps produced by the top-down saliency detector proposed in this work, when the tasks are, respectively, to detect cars and people. The main advantage of tying saliency to the recognition goal is that saliency judgements become significantly more adaptive, only highlighting image areas relevant for recognition. The same detector can produce significantly different saliency judgements for the same image, depending on what is to be detected.

In this work, we propose a **discriminant** principle for top-down saliency, denoted by *discriminant saliency*, which is *intrinsically connected to the recognition problem.* We start from the intuition that, for recognition, the *salient features of a visual class are those that best distinguish it from all other visual classes of recognition interest.* This intuition translates naturally into a *computational principle* for the design of top-down saliency detectors: classification with minimal expected probability of error. It is shown that this principle is closely related to a number of previously proposed principles for the organization of perceptual systems: maximization of information transmission across perceptual layers (infomax) [33]–[35], inference by detection of suspicious coincidences [36], and classification with minimal uncertainty. For visual saliency, these principles equate optimal features to those maximally informative of presence/absence of the target class in the field of view, whose observation is most suspicious in the absence of the target class, or which minimize the uncertainty about that presence/absence.

We investigate how these principles can be implemented with *computational parsimony,* by exploiting known properties of natural image statistics. It is shown that Barlow's principle of *inference by detection of suspicious coincidences* [36] enables computationally efficient saliency measures that are nearly optimal in the minimum probability of error sense. Barlow's principle is then used to derive computationally efficient algorithms for the two fundamental operations of discriminant saliency: *feature selection* and *saliency detection*. Experimental evaluation on object recognition tasks shows that the resulting top-down saliency detector can effectively act as a *focus-of-attention* mechanism, capable of pruning away bottom-up salient locations which are irrelevant for recognition. It is shown that this pruning improves the performance of state-of-the-art object recognition systems, in terms of both localization and classification accuracy. Finally, we show that discriminant saliency can adapt to a *rich* set of visual attributes.

The remainder of the paper is organized as follows. Section II introduces discriminant saliency and reviews its relations to previously proposed computational principles for perception. The implementation of feature selection and saliency detection under constraints

of computational parsimony is then discussed in Section III. The proposed saliency detector is introduced in Section IV, and an experimental evaluation of its performance discussed in Section V. Finally, some conclusions are drawn in Section VI.

II. Saliency principles

We start by introducing the concept of discriminant saliency, and relating it to previous principles for the organization of perceptual systems.

A. Discriminant saliency

Discriminant saliency is rooted in a decision-theoretic interpretation of perception. Under this interpretation, perceptual systems evolve to produce decisions about the state of the surrounding environment that are *optimal in a decision-theoretic sense*, e.g. that have minimum probability of error. To achieve this goal, discriminant saliency is defined with respect to two classes of stimuli: a *target class*, and a *null hypothesis*, composed of all the stimuli that are not salient. The locations of the visual field that can be classified, with greatest confidence, as containing target stimuli are denoted as salient. This definition of saliency is applicable to a broad set of problems. For example, different specifications for target stimuli and null hypothesis enable its specialization to both bottom-up and top-down saliency. We have previously studied discriminant saliency in the context of bottom-up saliency detection, by combining it with center-surround image processing, and shown that the resulting detector is biologically plausible and replicates various results from the psychophysics of human saliency [37], [38]. In this work, we consider the implementation of top-down discriminant saliency, and its benefits for recognition.

For this, we define top-down saliency detection with respect to a one-vs-all classification problem, where target stimuli are drawn from an object class of interest, and the null hypothesis is composed of the stimuli drawn from all other object classes that make up the recognition problem. Visual stimuli are not measured directly, but through their projection onto a set of basis functions, or *visual features*. These features can be seen as matched detectors to certain attributes of the visual stimulus. The features that best *discriminate between target and null hypotheses are deemed salient.* These are matched detectors to the salient attributes of the target. *Salient locations* are then defined as the locations of the visual field where the classification of the visual stimulus into target and non-target can be made with *highest confidence*.

Fig. 2.   Feature saliency depends on the viewing context.

When compared to bottom-up saliency, this definition has three interesting properties. First, by definition of target and null hypotheses, it makes saliency contingent upon the recognition problem, tuning it to the image attributes that best distinguish target from other object classes. Second, for a given object class, the saliency of a set of visual attributes changes with the recognition context. As the null hypothesis varies, so do the attributes that determine saliency. This is consistent with biological perception. For example, as illustrated by Figure II-A, when a white fox is viewed against a forest, its color is salient and recognition is easy. On the other hand, when the fox is presented against white snow, color is no longer a salient attribute and recognition becomes very difficult. Third, and perhaps most importantly, discriminant saliency translates easily into an optimality criterion for the design of saliency algorithms. This design consists of two steps. The first is an optimal feature selection problem: the identification of the visual features that optimally discriminate between target and null hypothesis. The second is an optimal decision-making problem: the identification of the locations, in the visual field, where the presence of these features can be most confidently attributed to the target class. Both feature discrimination and classification confidence can be quantified in various ways, most of which are directly connected to previously proposed principles for the organization of perceptual systems.

B. The minimum Bayes error principle

For a classification problem defined by a feature space $\mathcal{X}$, and a random variable Y that assigns $\mathbf{x} \in \mathcal{X}$ to one of M classes, $i \in \{1, \ldots, M\}$, the minimum probability of classification error is achieved by the *Bayes* classifier [39]

$$g^*(\mathbf{x}) = \arg\max_i P_{Y|\mathbf{X}}(i|\mathbf{x}). \tag{1}$$

This probability of error is denoted as the *Bayes error* (BE)

$$L^* = 1 - E_{\mathbf{x}}[\max_i P_{Y|\mathbf{X}}(i|\mathbf{x})], \tag{2}$$

where $E_{\mathbf{x}}$ is the expectation with respect to $P_{\mathbf{X}}(\mathbf{x})$. Since 1) the Bayes error depends only on $\mathcal{X}$, 2) lower bounds the probability of error of any classifier on $\mathcal{X}$, and 3) there is at least one classifier (the Bayes classifier) that achieves this lower bound, minimization of Bayes error is a natural optimality criteria for feature selection. Its minimization is, however, difficult due to the non-linearity of the $\max(\cdot)$ operator in (2).

To relate the minimization of Bayes error to other discriminant principles, we note that $P_{Y|\mathbf{X}}(i|\mathbf{x})$ can be interpreted as a *measure of confidence with which* $\mathbf{x}$ *can be assigned to class* i. Defining $c(\mathbf{x})$ as the *confidence measure* for the classification of $\mathbf{x}$, it follows that, under Bayes decision theory,

$$c^*(\mathbf{x}) = \max_i P_{Y|\mathbf{X}}(i|\mathbf{x}). \tag{3}$$

The (optimal) decision rule is then to select the class of highest confidence, and optimal feature selection corresponds to the choice of $\mathcal{X}$ which maximizes the expected confidence on the classification decisions, $E_{\mathbf{X}}[c(\mathbf{x})]$. Under Bayes decision theory this expected confidence is

$$C^* = E_{\mathbf{X}}[c^*(\mathbf{x})] = E_{\mathbf{X}}[\max_i P_{Y|\mathbf{X}}(i|\mathbf{x})], \tag{4}$$

and its maximization is equivalent to the minimization of the Bayes error.

C. Relaxation and the principle of minimum uncertainty

The analysis above suggests a procedure to obtain consistent decision rules and feature selection costs: define a confidence measure $c(\mathbf{x})$, a feature selection cost $E_{\mathbf{X}}[c(\mathbf{x})]$, and a decision rule which selects the class of highest confidence. One possibility to eliminate the non-linear $max(\cdot)$ operator inherent to the Bayes error, is to *relax* it, replacing $\max(\log p, \log(1-p))$ by $mean(\log p, \log(1-p)) = p \log p + (1-p) \log(1-p)$. We refer to this procedure as a *relaxation to the mean*. Noting that Bayes decision rule is identical to

$$g^{'}(\mathbf{x}) \quad = \quad \arg\max_i \log P_{Y|\mathbf{X}}(i|\mathbf{x}), \tag{5}$$

the application of relaxation (to the mean) to this rule leads to the confidence measure

$$c^{'}(\mathbf{x}) = -H(Y|\mathbf{X} = \mathbf{x}), \tag{6}$$

where $H(Y|\mathbf{X} = \mathbf{x}) = -\sum_i P_{Y|\mathbf{X}}(i|\mathbf{x}) \log P_{Y|\mathbf{X}}(i|\mathbf{x})$ is the entropy of the class label Y given the observation $\mathbf{X} = \mathbf{x}$. We say that $c^{'}(\mathbf{x})$ is consistent with Bayes decision rule, up to a relaxation to the mean. In this case, the expected confidence $E_{\mathbf{X}}[H(Y|\mathbf{X} = \mathbf{x})] = -H(Y|\mathbf{X})$ is the negative of the posterior entropy of Y given $\mathbf{X}$. Under this criterion, the optimal $\mathcal{X}$ is the one which minimizes the uncertainty of the classification decision, where uncertainty is measured in the standard information theoretic sense (entropy). It follows that uncertainty minimization is equivalent to the minimization of Bayes error, up to a relaxation to the mean.

D. Infomax and Barlow's principle of suspicious coincidences

Using the well known property that

$$I(Y;\mathbf{X}) = H(Y) - H(Y|\mathbf{X}), \tag{7}$$

where

$$I(\mathbf{X};Y) \;=\; \sum_i \int P_{\mathbf{X},Y}(\mathbf{x},i) \log \frac{P_{\mathbf{X},Y}(\mathbf{x},i)}{P_{\mathbf{X}}(\mathbf{x})P_Y(i)} d\mathbf{x} \tag{8}$$

is the mutual information between class label Y and feature vector $\mathbf{X}$ and

$$H(Y) = -\sum_i P_Y(i) \log P_Y(i)$$

the class entropy, and the fact that $H(Y)$ does not depend on X, it follows that uncertainty minimization is equivalent to selecting the features that have largest mutual information with the class label [40]–[43]. This is frequently referred to as the *infomax* criteria, due to its connections to the *infomax* principle for the organization of perceptual systems, a principle of long traditions in cognitive science [33]–[35]. The underlying confidence measure

$$c^{''}(\mathbf{x}) = I(Y;\mathbf{X}=\mathbf{x}) = \sum_i P_{Y|\mathbf{X}}(i|\mathbf{x}) \log \frac{P_{\mathbf{X},Y}(\mathbf{x},i)}{P_{\mathbf{X}}(\mathbf{x})P_Y(i)}, \tag{9}$$

is a relaxation (to the mean) of the decision rule

$$g^{''}(\mathbf{x}) \;=\; \arg\max_i \log \frac{P_{Y,\mathbf{X}}(i,\mathbf{x})}{P_Y(i)P_{\mathbf{X}}(\mathbf{x})}. \tag{10}$$

This decision rule was proposed by Barlow as the fundamental computation of cerebral cortex [36]. He argued that a nerve cell "represents a hypothesis about the sense organs it connects with" and "the multitude of nerve cells in sensory pathways constantly test a multitude of hypotheses about the *environment*." He then equated the cortex to a detective that "makes inductive inferences about the environment" by "looking out for *suspicious coincidences*. The occurrence of two events A and B is suspicious if they occur jointly more often than what would be expected from the probabilities of individual occurrence, i.e. $P(A,B) >> P(A)P(B)$." "The

TABLE I

DECISION RULES, CONFIDENCE MEASURES, AND FEATURE SELECTION COSTS.

	rule	confidence	cost	principle
Bayes	$g^*(\mathbf{x})$	$c^*(\mathbf{x})$	$-L^*$	Min. error probability
Bayes (relaxed)	$g^{'}(\mathbf{x})$	$-H(Y\|\mathbf{X}=\mathbf{x})$	$-H(Y\|\mathbf{X})$	Min. uncertainty
Barlow (relaxed)	$g^{''}(\mathbf{x})$	$I(Y;\mathbf{X}=\mathbf{x})$	$I(Y;\mathbf{X})$	Infomax
parsimony	$g_k^{''}(x)$	$\sum_k I(Y;X_k=x)$	$\sum_k I(Y;X_k)$	Infomax

cortical neurons in a region share amongst themselves the task of detecting the coincidences that occur on the input fibers of that region.” In computer vision, the detection of suspicious, or non-accidental, coincidences has been proposed as a principle for perceptual organization by various authors [44], [45].

It follows from the equivalence between infomax and uncertainty minimization that infomax is equivalent to the minimization of Bayes error, up to a relaxation to the mean. The different combinations of decision rule, confidence measure, and feature selection cost are summarized in Table I. It is quite interesting that, although the decision rules of Bayes and Barlow are different (one minimizes error probability, the other seeks maximally suspicious coincidences), their relaxation produces identical feature selection costs (up to the constant $H(Y)$ which does not depend on $\mathbf{x}$). In this sense, they are *identical for feature selection*.

III. LEARNING SALIENT FEATURES

Since the goal of top-down saliency is to identify regions for further processing, it should be computationally efficient. This implies that the selection of optimally discriminant features should itself be subject to constraints of computational parsimony. We investigate how to enforce such constraints by exploiting statistical properties of natural image features.

A. Natural image statistics and computational efficiency

One appealing property of infomax feature selection, and a substantial advantage over minimum Bayes error, is its potential for computational parsimony. Defining $\mathbf{X}_{1,k} = (X_1, \ldots, X_k)$, (8) can be rewritten as

$$I(Y;\mathbf{X}) = \sum_k I(Y;X_k) + \sum_k [I(X_k;\mathbf{X}_{1,k-1}|Y) - I(X_k;\mathbf{X}_{1,k-1})] \tag{11}$$

where

$$I(\mathbf{X};Y|\mathbf{Z}) \;=\; \sum_i \int P_{\mathbf{X},Y,\mathbf{Z}}(\mathbf{x},i,\mathbf{z}) \log \frac{P_{\mathbf{X},Y|\mathbf{Z}}(\mathbf{x},i|\mathbf{z})}{p_{\mathbf{X}|\mathbf{Z}}(\mathbf{x}|\mathbf{z})p_{Y|\mathbf{Z}}(i|\mathbf{z})} d\mathbf{x}d\mathbf{z}. \tag{12}$$

The terms $I(X_k;Y)$ quantify the discriminant information conveyed by each feature, and the terms $I(X_k;\mathbf{X}_{1,k-1}|Y)-I(X_k;\mathbf{X}_{1,k-1})$ quantify the discriminant information contained in feature dependencies [43]. This can be combined with Attneave's hypothesis [34] that perception is tuned to the environment, to achieve substantial reductions in complexity. Of particular interest is a known statistical property of band-pass features, such as wavelet coefficients, extracted from natural images: that such features exhibit strongly *consistent* patterns of dependency across a very wide range of natural image classes [46], [47]. For example, when a natural image is subject to a wavelet decomposition, the conditional distribution of any wavelet coefficient, given the state of the co-located coefficient of immediately coarser scale (known as its "parent"), invariably has a bow-tie shape [46]. It follows that, while the coefficients are statistically dependent, their dependencies carry little information about the image class [43], [46]. Hence, the second summation of (11) is much smaller than the first, and (8) is well approximated by

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

Note that this approximation *does not* assume that the features are independently distributed, but simply that their dependencies are not informative about the class. This is a new feature selection cost, and the expectation of a new confidence measure

$$c^{'''}(\mathbf{x}) = \sum_k I(Y;X_k = x_k), \tag{14}$$

which results from relaxing (to the mean) the decision rules

$$g_k^{''}(x) = \arg\max_i \log \frac{P_{Y,X_k}(i,x)}{P_{X_k}(x)P_Y(i)}, k \in \{1,\ldots,K\}. \tag{15}$$

Note that this is a set of decision rules which act on the feature channels individually. We refer to this type of independent application of a decision rule to each channel as a *marginal decision rule*. The adoption of (13) enables two substantial simplifications. First, because mutual information is always positive, a very simple feature selection strategy is globally optimal when (13) holds [41]: to select the K optimal features it suffices to 1) order all features by decreasing $I(X_k,Y)$, and 2) select the first K. Second, the terms in the right hand side of (13) only require marginal density estimates. As we will see in the next section, these are extremely simple for bandpass features extracted from natural images.

B. Computational parsimony and suspicious coincidences

Returning to table I, there are two important points to note. The first is that there is no equivalent to (11) for the feature selection cost of (4). Due to this, although it would be possible to define a computationally parsimonious feature selection cost of the form

$$C = \sum_k E_{X_k}\left[\max_i P_{Y|X_k}(i|x_k)\right]$$

it does not necessarily follow that such a cost would be a good approximation to (4). In fact, the relaxation of the max appears to be a necessary condition for the conjunction of near optimality and computational parsimony. Whether this relaxation has to be to the mean is, at this point, not known. Second, while the relaxation of (15) is sufficient for parsimony, the latter does not necessarily hold for relaxations of all marginal decision rules. In particular, even though the maximization of $-H(Y|\mathbf{X})$ and $I(Y;\mathbf{X})$ *both* lead to the infomax solution for feature selection, (13) does not imply that $H(Y|\mathbf{X}) \approx \sum_k H(Y|X_k)$. Instead, it can be shown, by application of (7), that it is identical to $H(Y|\mathbf{X}) \approx \sum_k H(Y|X_k) - (K-1)H(Y)$. The confidence measure associated with this approximation *is still* (14) and *not* the relaxation of the marginal Bayes decision rules

$$g_k^*(x) = \arg\max_i \log P_{Y|X_k}(i|x), k \in \{1, \dots, K\}. \quad (16)$$

In this way, the constraint of computational parsimony *breaks* the connection between infomax feature selection and relaxations of Bayes decision rule. In fact, among all decision rules considered so far, only the marginal rules $g_k^{''}(x)$ of (15) are consistent (up to a relaxation to the mean) with (13). For this reason, we believe that the detection of suspicious coincidences is preferable to the explicit minimization of error probability when there are constraints of computational parsimony, and adopt this principle in the remainder of the work.

Note that, at this point, we do not have a decision rule for the observed feature vector $\mathbf{x}$, but the collection of marginal decisions of (15). This is intuitive: in the absence of discriminant feature dependencies, the detection of globally suspicious coincidences simplifies into the detection of suspicious coincidences in each feature channel. It is also consistent with the psychophysics of human saliency, where it is well known that humans can easily detect differences between targets and distractors along a single dimension (e.g. different color), but not when they require a conjunction of features (e.g. differences in the conjunction of color and orientation) [48]. Finally, the absence of a holistic decision rule for $\mathbf{x}$ is not problematic for saliency. In general, search tasks are better served by a graded measure of confidence $c(\mathbf{x})$ on how salient each feature vector $\mathbf{x}$ is, than by a hard binary classification. When considering multiple salient locations, attention should first be deployed to the location that can be declared salient with greatest confidence. If the object is not found there, the location of the next largest confidence should be inspected, and so forth. This is the rationale behind all existing saliency detectors, which search through local maxima of some saliency function [4], [5]. It is also consistent with the mechanisms of inhibition of return commonly found in biological vision [3], [49]. A holistic confidence measure is well defined for the principle of suspicious coincidences, under parsimony constraints: it consists of the sum of marginal confidence measures of (14).

C. The generalized Gaussian distribution

Besides (13), a well known property of the statistics of natural images can be exploited to increase computational efficiency: that the probability densities of band-pass image features extracted from such images are well approximated by a generalized Gaussian distribution (GGD) [47], [50], [51],

$$P_X(x;\alpha,\beta) = \frac{\beta}{2\alpha\Gamma(1/\beta)} \exp\left\{-\left(\frac{|x|}{\alpha}\right)^{\beta}\right\}, \tag{17}$$

where $\Gamma(z) = \int_0^\infty e^{-t}t^{z-1}\mathrm{d}t$, $t > 0$, is the Gamma function, α is a *scale* parameter, and β is a *shape* parameter. The parameter β controls the rate of decay from the peak value, and defines a sub-family of the GGD (e.g. Laplacian when $\beta = 1$ or Gaussian when $\beta = 2$). The GGD has various interesting properties. First, various low-complexity methods exist for the estimation of the parameters (α, β), including the method of moments [52], maximum likelihood [53] and minimum mean-square-error [47]. We adopt the method of moments in what follows. The two parameters are estimated from

$$\sigma^2 = \frac{\alpha^2\Gamma(\frac{3}{\beta})}{\Gamma(\frac{1}{\beta})} \quad \text{and} \quad \kappa = \frac{\Gamma(\frac{1}{\beta})\Gamma(\frac{5}{\beta})}{\Gamma^2(\frac{3}{\beta})}, \tag{18}$$

where σ^2 and κ are, respectively, the variance and kurtosis of X

$$\sigma^2 = E_X[(X - E_X[X])^2], \text{ and } \kappa = \frac{E_X[(X - E_X[X])^4]}{\sigma^4}.$$

This method has been shown to produce good fits to natural images [47].

Second, it leads to closed form solutions for various information theoretic quantities. For example, when both $P_{X|Y}(x|i)$ and $P_X(x)$ are well approximated by GGDs, the mutual information

$I(X;Y)$ has a closed form. This follows from

$$I(\mathbf{X};Y) \;=\; \sum_i P_Y(i) KL\left[P_{\mathbf{X}|Y}(\mathbf{x}|i)||P_{\mathbf{X}}(\mathbf{x})\right] \tag{19}$$

where $KL[p||q] = \int p(\mathbf{x}) \log \frac{p(\mathbf{x})}{q(\mathbf{x})} d\mathbf{x}$ is the Kullback-Leibler (KL) divergence between $p(\mathbf{x})$ and $q(\mathbf{x})$, and [53],

$$KL[P_X(x;\alpha_1,\beta_1)||P_X(x;\alpha_2,\beta_2)] = \log\left(\frac{\beta_1\alpha_2\Gamma(1/\beta_2)}{\beta_2\alpha_1\Gamma(1/\beta_1)}\right) + \left(\frac{\alpha_1}{\alpha_2}\right)^{\beta_2} \frac{\Gamma((\beta_2+1)/\beta_1)}{\Gamma(1/\beta_1)} - \frac{1}{\beta_1}. \tag{20}$$

It can also be shown that

$$H(X|Y=i) = \frac{1}{\beta_i} + \log \frac{2\alpha_i\Gamma(\frac{1}{\beta_i})}{\beta_i} \tag{21}$$

and

$$I(Y;X=x) = s[g(x)] \log \frac{s[g(x)]}{\pi_1} + s[-g(x)] \log \frac{s[-g(x)]}{\pi_0}, \tag{22}$$

where $s(x) = (1+e^{-x})^{-1}$ is a sigmoid function, $\pi_i = P_Y(i)$ is the prior for class i, and $g(x) = \left(\frac{|x|}{\alpha_0}\right)^{\beta_0} - \left(\frac{|x|}{\alpha_1}\right)^{\beta_1} + T$, with $T = \log \frac{\alpha_0\beta_1\pi_1\Gamma(1/\beta_0)}{\alpha_1\beta_0\pi_0\Gamma(1/\beta_1)}$. We will make use of these closed forms to derive an efficient implementation of discriminant saliency.

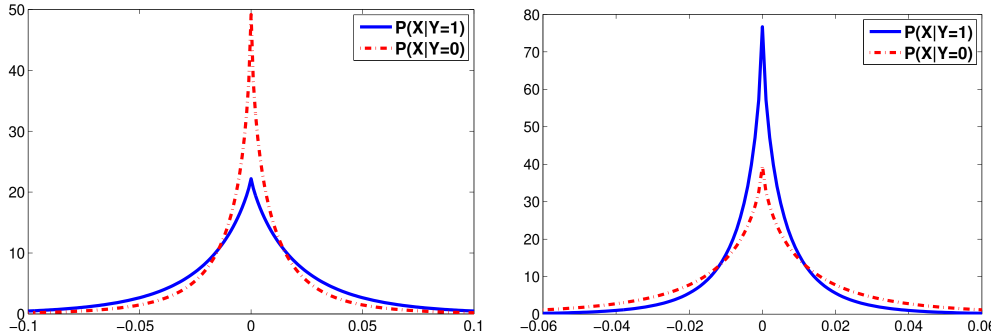
IV. Discriminant saliency

The design of a discriminant saliency detector has two components: feature selection and saliency detection.

A. *Feature selection*

We have seen in Section III-A that, given a space $\mathcal{X}$ of band-pass features extracted from natural images, the best K-feature subset can be selected by computing the marginal mutual informations $M_k = I(Y;X_k)$, for all k, and selecting the K features of largest M_k. The marginal mutual informations can be computed efficiently with (19) and (20). One final issue is that none of the feature selection costs considered so far is asymmetric: in general, discrimination does not differentiate between situations where 1) the feature is present (strong responses) in the object class of interest, but absent (weak response) in the null hypothesis, and 2) vice versa. Although both cases lead to low probability of error, feature absence is less interesting for saliency, which is an inherently asymmetric problem.

However, detecting if a feature is discriminant due to presence or absence in the class of interest is usually not difficult. For generalized Gaussian features, it suffices to note that feature



(a) (b)

Fig. 3. Illustration of the conditional marginal distributions (GGDs) for the responses of a feature, when (a) it is present (strong responses) in the object class ($Y = 1$) but absent (weak responses) in the null hypothesis ($Y = 0$), or (b) vice versa. Note that the absence of a feature always leads to narrower GGDs than the presence of the feature.

absence produces a narrow GGD, close to a delta function, while feature presence increases the variance of the distribution (see Figure 3 for an example). Since a narrow GGD has lower entropy than one of larger variance, discriminant features which are absent from the class of interest fail the test

$$H(X_k|Y = 1) > H(X_k|Y = 0), \tag{23}$$

or, using (21),

$$\log \frac{\alpha_1}{\alpha_0} > \left(\frac{1}{\beta_0} - \frac{1}{\beta_1}\right) + \log \frac{\Gamma(\frac{1}{\beta_0})\beta_1}{\Gamma(\frac{1}{\beta_1})\beta_0}. \tag{24}$$

Such features should not be considered during feature selection.

B. Saliency detection

Under discriminant saliency, all saliency judgments are based on a measure of classification confidence $c(\mathbf{x})$. As before, the asymmetric nature of the saliency problem needs to be taken into account: feature vectors that can be *very confidently* classified as *not* belonging to the class of interest should *not* be declared salient. This is accomplished by introducing a decision rule which summarily eliminates such feature vectors. Both the confidence measure and the decision rule should be consistent with (13). The fact that, in Table I, only (14) and (15) satisfy this requirement, leads to the saliency measure

$$S_D(\mathbf{x}) = \sum_{k=1}^{K} S_k(x_k), \tag{25}$$

with

$$S_k(x) = \begin{cases} I(Y; X_k = x) & \text{if } x \in \mathcal{S}_k \\ 0, & \text{otherwise,} \end{cases} \tag{26}$$

and

$$\mathcal{S}_k = \left\{ x \,\middle|\, \frac{P_{Y,X_k}(1,x)}{P_Y(1)P_{X_k}(x)} > \frac{P_{Y,X_k}(0,x)}{P_Y(0)P_{X_k}(x)} \right\}. \tag{27}$$

This measure has various interesting properties. First, it implements Barlow's principle of suspicious coincidences by 1) identifying features whose appearance in the field of view is suspiciously coincident with that of the object class of interest, (27), and 2) equating saliency to the associated (log) degree of suspicion

$$S_k(x) = \left\langle \log \frac{P_{Y,X_k}(i,x)}{P_Y(i)P_{X_k}(x)} \right\rangle,$$

where $\langle f(x) \rangle = \sum_i P_{Y|X}(i|x) f(x)$. The overall saliency measure $S_D(\mathbf{x})$ is the cumulative degree of suspicion over all feature channels.

Second, it equates salient features to *matched filters* for the detection of the salient visual attributes of the object of interest. This follows from the facts that 1) $\mathcal{S}_k$ can be written as

$$\mathcal{S}_k = \left\{ x \,\middle|\, P_{X_k|Y}(x|1) > P_{X_k|Y}(x|0) \right\}, \tag{28}$$

and 2) GGD features which pass the test of (23) have a narrower $P_{X_k|Y}(x_k|0)$ than $P_{X_k|Y}(x_k|1)$ (see Figure 3). In result, $\mathcal{S}_k$ is of the form $\mathcal{S}_k = \{x \,||x| > t_k\}$, where t_k is a threshold that depends on the parameters of the two GGDs, and only regions of large magnitude feature response are salient. This implies that the features are matched to the visual stimuli considered salient.

Finally, it has an intuitive interpretation as a mechanism for the *allocation of attention*. This follows from rewriting $\mathcal{S}_k$ as $\mathcal{S}_k = \left\{ x \,\middle|\, P_{Y|X_k}(1|x_k) > P_Y(1) \right\}$. At first analysis, when compared with Bayes decision rule ($P_{Y|X_k}(1|x_k) > 1/2$), this appears sub-optimal for low-probability objects. It should, however, be noted that Bayes decision theory is broader than we have considered so far. While a threshold of $1/2$ minimizes the expected probability of error, this minimum is of interest only when false-positives (spurious detections) and false-negatives (undetected targets) have equal costs. When this is not the case, the threshold is determined by the ratio of the costs. In summary, there is no single "optimal" threshold: different thresholds are optimal under different cost structures. A threshold of $P_Y(1)$ ties the cost structure to the observer's prior beliefs on target likelihood, making the observer more or less conservative according to these beliefs. Searches for very unlikely targets have a low threshold, and require the inspection of a large number of locations. On the other hand, many locations are summarily rejected in searches for very likely targets. If the number of inspected locations is interpreted as the *amount of attention devoted to the scene*, this is an intuitive search behavior: searches for *rare* targets require *more attention* than searches for *frequent* ones. It motivates the use of (25) as a *focus-of-attention* mechanism.

C. Discriminant saliency as a focus-of-attention mechanism

In biology, visual saliency usually combines bottom-up and top-down saliency mechanisms (e.g. [4]). Top-down saliency can be seen as a focus-of-attention mechanism, which prunes away bottom-up salient points in regions unlikely to contain objects of interest. We have implemented this focus-of-attention strategy by combining discriminant saliency with classical bottom-up operators from computer vision. The details of the resulting top-down interest point detector are given in Algorithm 1. Locations of bottom-up saliency are first identified with classic interest point detectors. Top-down saliency, with respect to the object class of interest, is then computed with (25). The two saliency channels are combined by weighing interest points according to the value of discriminant saliency at their center location. The weighted points are finally sorted by decreasing top-down saliency, and used for either object localization or image classification, as discussed in the following section. While pruning interest points is not the only application of discriminant saliency, it enables an objective evaluation of the benefits of discriminant saliency for computer vision. In particular, because interest points are commonly used in state-of-the-art object recognition systems, it suffices to measure how the localization/classification accuracy of the latter varies with the addition of top-down attentional pruning.

V. Experimental evaluation

The performance of the proposed discriminant saliency detector (DSD) was evaluated on a set of weakly supervised object recognition experiments.

A. Experimental setup

1) Datasets: Weakly supervised object recognition addresses the design of object recognition systems from informally collected examples. In particular, training examples of the object of interest are presented against cluttered backgrounds. Two tasks are usually considered. The first is

Algorithm 1 Top-down interest point detection

Training: Given a set of features $X_k, k \in \{1, \ldots, N\}$, a set of images $\mathcal{T}_1$ from the target class, a set of images $\mathcal{T}_0$ from the null hypothesis, and a target number of features K.

for $k = \{1, \ldots, N\}$ **do**

Estimate GGD parameters of $P_{X_k|Y}(x|i)$, from responses of X_k to $\mathcal{T}_i$, $i \in \{0, 1\}$, using (18).

Check whether X_k passes the test of (24). If not, discard X_k and move to feature $k + 1$.

Estimate GGD parameters of $P_{X_k}(x)$, from responses of X_k to $\mathcal{T}_0 \cup \mathcal{T}_1$, using (18).

Compute $I(X_k, Y)$, using (19) and (20).

end for

Output: return the K features of largest $I(X_k, Y)$.

Saliency detection: Given a test image $\mathcal{I}$, a set of K discriminant features X_k for the target class, and the GGD parameters of $P_{X_k|Y}(x|i)$, $i \in \{0, 1\}$, and $P_{X_k}(x)$.

Determine a set of interest point locations $\mathbf{l}_1, \ldots, \mathbf{l}_M$, using standard interest point operators.

for $k = \{1, \ldots, K\}$ **do**

for $m = \{1, \ldots, M\}$ **do**

Compute the response x_m of X_k at location $\mathbf{l}_m$ of $\mathcal{I}$, and $P_{X_k|Y}(x_m|i)$, $i \in \{0, 1\}$, using (17).

Compute $S_k(x_m)$, using (26), (28), and (22).

end for

end for

for $m = \{1, \ldots, M\}$ **do**

Compute the saliency value at $\mathbf{l}_m$ with (25).

end for

Output: Return a list of interest points $\mathbf{l}_m$ ordered by decreasing discriminant saliency values $S_D(\mathbf{x}_m)$.

to determine whether a given image contains an instance of the object of interest (*classification*). The second is to locate the image area covered by the latter (*localization*). For both tasks we adopt the *PASCAL2006* dataset [54], which contains 10 object categories: bicycle, bus, car, cat, cow, dog, horse, motorbike, person, and sheep. It is an interesting dataset because the images were collected with limited control over appearance and pose of objects and background, and many images contain instances of several classes. It also provides ground truth for each object, since all images are annotated with bounding boxes. Although the ground truth was not used for training, its availability allows objective measurements of localization accuracy. **The images were divided into training and test sets, according to [54], and the numbers of images in each set are listed in Table II.** For each category, we defined a one-vs-all classification problem opposing the object class under consideration to the others.

2) Bottom-up interest points: Bottom-up interest points are extracted with three operators widely used in object recognition: Harris-Laplace (HarrLap) [26], Hessian-Laplace (HesLap) [26], and difference of Gausssian (DoG) [24]. The binaries are available from `http://lear.`

TABLE II

TRAINING AND TESTING IMAGES OF PASCAL2006.

	bicycle	bus	car	cat	cow	dog	horse	motorbike	person	sheep
# of Training	270	174	553	386	206	365	247	235	666	251
# of Testing	268	180	544	388	197	370	254	234	675	238

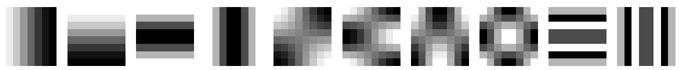


Fig. 4. Some of the basis functions in the DCT feature set.

`inrialpes.fr/people/dorko/downloads.html`.

3) Candidate features for discriminant saliency: The discriminant saliency detector does not have free parameters. The only component left to be specified is the initial pool of band-pass features. We have obtained very similar results with Gabor filters, Haar wavelets, and the discrete cosine transform (DCT) [55]. The implementation discussed here is based on a *multi-scale extension of the DCT*. Each image is decomposed into a five-level Gaussian pyramid, and 8×8 DCT features computed at each location and scale, **for a total of** 320 **features.** The so-called DC coefficient (average of the image patch) is discarded at all scales, to guarantee lighting invariance. As shown in Figure 4, many of the DCT basis functions can be interpreted as detectors of perceptually relevant image attributes, including edges, corners, t-junctions, and spots. **The number of features used by the discriminant saliency detector for each object class (**K **in Algorithm 1) is the number of features that pass the test of (24).**

4) Preliminary analysis of contextual influences: It is well known that background scenes provide a substantial amount of *contextual* information, which can significantly simplify the recognition task [13], [56], [57]. While, in general, the ability to exploit context is an asset, it is not the goal of saliency, which aims to identify the objects themselves. This makes it important to verify that a saliency detector performs well because it has learned to recognize the objects of interest, and not the background. In PASCAL2006, context is a strong cue for the recognition of some object classes. For example, cows and sheep always stand on grass, while cars and buses are surrounded by roads, buildings, or parking lots. To quantify the strength of these contextual cues, we performed a preliminary image classification experiment.

The classifier was that proposed by [13], and shown to achieve state-of-the-art results on PASCAL2006. It combines the "bag-of-features" image representation with the spatial

pyramid matching technique (and is referred to as the *BFSP* classifier). Both its implementation and the selection of all parameters followed closely [13]. The bag-of-features representation is based on a dictionary of visual words, which are cluster centers of SIFT descriptors extracted at interest points locations from all training images. K-means was used for clustering and the dictionary contains 3000 visual words. The spatial pyramid representation is obtained by 1) repeatedly subdividing the image into increasingly finer sub-regions, and 2) characterizing each sub-region by a histogram of the visual words found inside it [31]. This representation is combined with a χ^2 kernel to train an SVM for each object class. Two classifiers were implemented. They were identical in everything except the interest point locations: the first used the bottom-up interest points located inside the ground truth object bounding box (referred to as *BU-GT*). The second used those located outside the box (referred to as *BU-BG*).

Table III presents the classification accuracy obtained for all classes, as **measured by the *Area Under ROC Curve (AUC)***. It is clear that, in some cases, the contextual cues are quite strong. For example, cow classification using merely background points (no object information) is only 3.56% less accurate than that **based uniquely on points located within the groundtruth bounding box for the object of interest**. Other classes with strong contextual influences are car, sheep, and bus. These contextual influences create two problems for saliency. First, a saliency detector can only learn to separate object from background if the background scenes are diverse [58]. If a cow always appears in a patch of grass, it is impossible to learn that cow and grass are different visual concepts. For discriminant saliency, this translates into the selection of features (e.g. grass descriptors) which are discriminant because they enable the detection of the "object-attached" background (grass), not the object itself. Note that this is not a problem for classification (assuming, of course, that the test images show cows in the same grass patch), but it is a problem for localization. In this sense, localization is a better measure of saliency detection performance than image classification. Second, a feature can be discriminant simply because it is consistently absent from the "object-attached" background. This is not a problem for discriminant saliency, since such features are rejected by the test of (23), unless the effect is overwhelming. For example, the ubiquity of large smooth regions (grass) in the cow class makes any textured feature discriminant for its detection. Since the animal body is also textureless, none of these features is informative for the class of interest, and all features are

TABLE III

IMAGE CLASSIFICATION ACCURACY (%) OF BFSP CLASSIFIERS FOR PASCAL2006.

	bicycle	bus	car	cat	cow	dog	horse	motorbike	person	sheep
BU-GT	96.18	99.34	97.89	94.32	92.64	85.78	92.59	98.56	87.42	95.39
BU-BG	81.62	91.63	92.51	86.25	89.08	76.20	81.99	82.72	73.20	89.13
Diff. (%)	14.56	7.71	5.38	8.07	3.56	9.58	10.60	15.84	14.22	6.26

discarded. In PASCAL2006, we observed this behaviour for the cow and sheep classes. The fact that discriminant saliency rejects all features is actually one of its advantages: it indicates that the underlying feature pool is not rich enough to account for these objects. In this case, it should contain shape features, which are important cues for cow or sheep detection. While we could have pursued an expanded feature set, we felt that this issue is complementary to the discussion of the paper, and eliminated the cow and sheep classes from further consideration.

B. Object localization

We next evaluate the localization accuracy of the combination of bottom-up interest point detection and top-down pruning, based on DSD. **Note that the goal is not to design a full-fledged object detection system, but simply to investigate whether salient points are co-located with the objects of interest. This is accomplished by measuring the percentage of salient points that land inside the groundtruth bounding boxes available in PASCAL. These experiments only consider images that contain the object of interest, and test the ability of the saliency detector to behave as a focus of attention mechanism tuned for object recognition.** We refer to the locations deemed salient by the combination of the bottom-up and top-down mechanisms as *DSD-BU points*, and **compare their localization to those of various discriminant interest point pruning strategies proposed in recent years [11], [12], [14], [16], [17], [59], [60]. In particular, three representative methods are selected for comparison.**

1) DVW, LSVM, and pLSA: The *discriminative visual words (DVW)* detector [14], [59] extracts interest points from a set of training images, and describes them by SIFT descriptors [24]. It then estimates the distribution of these descriptors with a Gaussian mixture model (GMM) and standard clustering techniques. Each cluster center is referred to as a *visual word*. Discriminant visual words are found with an estimate of the posterior probability of object i given word w

$$D(w) = P_{Y|W}(i|w) = \frac{\text{\# of times that } w \text{ appeared in images from class } i}{\text{\# of times that } w \text{ appeared in any image}}, \tag{29}$$

referred to as the discriminability of word w. The discriminability of an interest point is quantified by the posterior probability of the object given the SIFT descriptor ($\mathbf{x}$) at the point

$$P_{Y|\mathbf{X}}(i|\mathbf{x}) = \sum_w D(w) P_{W|\mathbf{X}}(w|\mathbf{x}), \qquad P_{W|\mathbf{X}}(w|\mathbf{x}) = \frac{P_{\mathbf{X}|W}(\mathbf{x}|w)P_W(w)}{\sum_w P_{\mathbf{X}|W}(\mathbf{x}|w)P_w(w)},$$

where $P_W(w)$, and $P_{\mathbf{X}|W}(\mathbf{x}|w)$ are the GMM components. **We implemented a DVW detector with the parameters of [14], using 3000 visual words learned with K-means.**

Like the DVW detector, the *linear support vector machine (LSVM)* selects discriminant visual words [60]. However, discriminability is measured with a linear SVM which classifies histograms of visual words. Following [61], the discriminability of a feature (or visual word) is equated to the absolute value of the weight given to that word by the SVM. To account for feature discrimination due to presence or absence in the object class, the discriminability of visual words which appear more often in the null hypothesis than in the target class is set to 0, akin to the test of (23). During detection, the saliency of an interest point is measured by the discriminability of the corresponding visual word. **We implemented the detector according to [60], using SIFT feature descriptors and 3000 visual words learned with K-means.**

Probabilistic Latent Semantic Analysis (pLSA) is a topic discovery method developed in the text analysis literature, and successfully applied to object categorization [11], [12]. It models each image as a document of visual words **(vectors of SIFT features)**. The set of images is equated to a text corpus, and pLSA learns object categories as mixtures of representative topics for this corpus. Mathematically, an image (or document d) is represented as a “bag” of visual words (w) sampled from a hidden topic variable (Z). The joint distribution is

$$P_{W,D,Z}(w,d,z) = P_D(d)P_{Z|D}(z|d)P_{W|Z}(w|z), \tag{30}$$

where $P_{W|Z}(w|z)$ and $P_{Z|D}(z|d)$ are multinomial distributions. Estimation of the pLSA model involves determining topic vectors representative of all documents, and the coefficients of each document, i.e. finding topic-specific word-distributions $P_{W|Z}(w|z)$ and document-specific mixing proportions $P_{Z|D}(z|d)$. This is done with the expectation-maximization (EM) algorithm [11], [62]. The pLSA model was originally proposed for unsupervised learning, but it was shown that, for cluttered data, the learned object topics are likely to include words representative of background clutter [11]. To avoid this, we train the model in the discriminant manner suggested in [11]. A set of “background” topics is first learned from negative training images, and frozen.

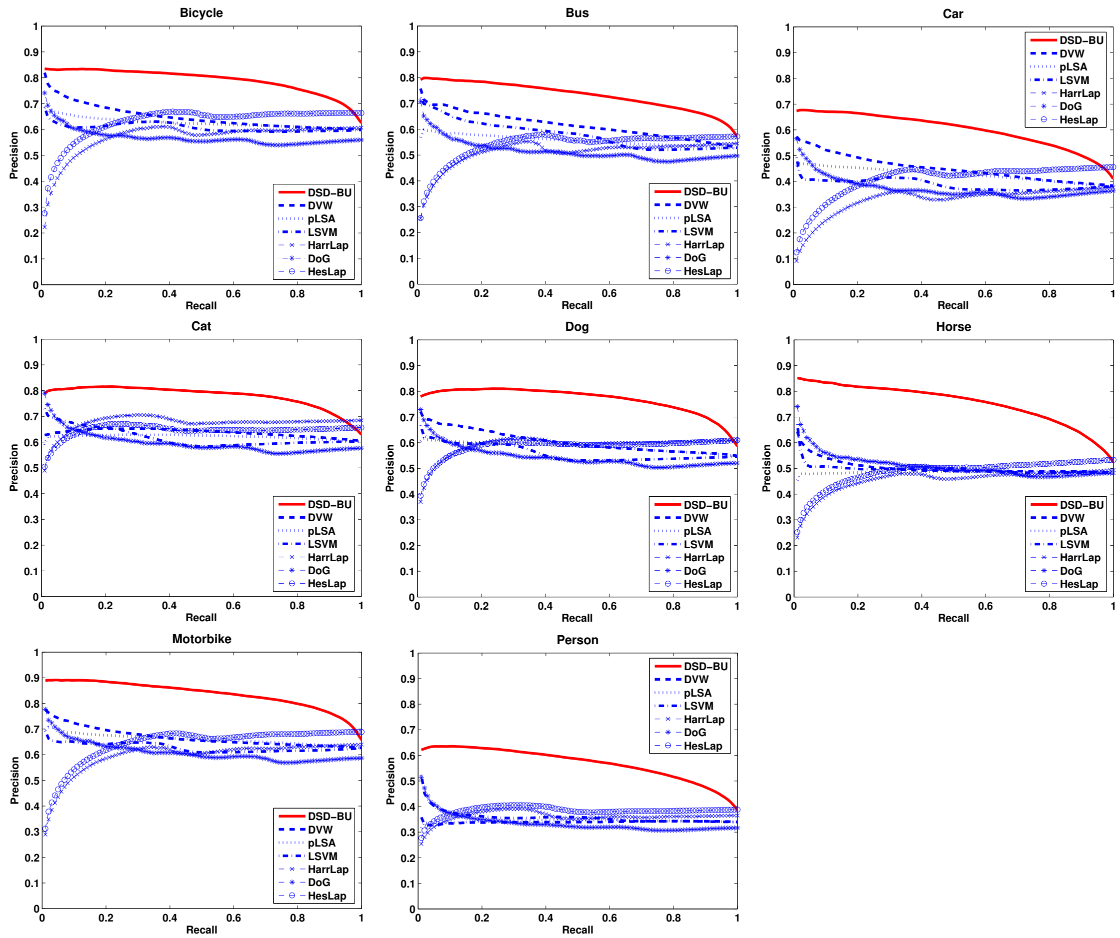


Fig. 5. Precision-recall of various saliency detectors on PASCAL2006.

The topics representative of each object class are then learned from the set of images of that class. One pLSA model is learned for each object category. For the detection of an object category, bottom-up interest points are pruned by the topic posterior probabilities, $P(z|w,d)$, at these points. Similarly to DVW and LSVM, the visual word dictionary consisted of 3000 words learned with K-means. The number of object and background topics were determined by cross-validation. Best results were achieved with seven background and one object topic. This was the configuration adopted to produce the results reported. **The implementation of pLSA was obtained from `http://www.robots.ox.ac.uk/~vgg/software/pLSA`.**

2) Localization accuracy: The localization accuracy of salient points was measured with precision/recall (PR) curves. Given a threshold, the points of saliency above it are classified as "object" and the remainder as "background". Denoting the ground-truth bounding box by $\mathcal{B}$, the following quantities are computed for each object/background assignment:

TABLE IV

AVERAGE PRECISION-RECALL FOR VARIOUS DETECTORS.

	bicycle	bus	car	cat	dog	horse	motorbike	person	median
DSD-BU	**0.76**	**0.72**	**0.60**	**0.78**	**0.76**	**0.75**	**0.82**	**0.56**	**0.76**
DVW	0.65	0.61	0.45	0.64	0.60	0.50	0.67	0.34	0.61
LSVM	0.59	0.58	0.37	0.62	0.56	0.49	0.62	0.35	0.57
pLSA	0.63	0.56	0.43	0.62	0.58	0.49	0.66	0.34	0.57
HarrLap	0.57	0.51	0.33	0.67	0.59	0.46	0.60	0.36	0.54
HesLap	0.63	0.54	0.41	0.65	0.58	0.49	0.64	0.38	0.56
DoG	0.57	0.52	0.37	0.60	0.54	0.51	0.61	0.33	0.53
Features	164	159	152	94	101	158	160	150	158

- True Positive (TP) - number of "object" points with center inside $\mathcal{B}$;
- False Negative (FN) - number of "background" points with center inside $\mathcal{B}$;
- False Positive (FP) - number of "object" points with center outside $\mathcal{B}$.

The recall and precision rates are then defined as

$$\text{Recall} = \frac{TP}{TP+FN} \qquad\qquad \text{Precision} = \frac{TP}{TP+FP}. \tag{31}$$

Localization accuracy is quantified as the average precision (AP) over the range of recalls.

3) Results and discussion: Figure 5 presents the PR curves of various detectors on PASCAL2006. A plot is presented for each object category, comparing the PR curves of the four top-down saliency detectors (DSD-BU, DVW, LSVM, and pLSA) and the three bottom-up interest point operators (HarrLap, HesLap, and DoG). **These plots can be most easily understood by noting that the value of precision at 100% recall is roughly equal to the average percentage of the image area occupied by the object of interest. This value of precision can thus be thought of as chance level performance: randomly selected points will fall within the object bounding box with this precision, at all levels of recall. It follows that a detector which is unable to discriminate between object of interest and background has a roughly constant PR curve, with precision equal to this value. On the other hand, a detector with good localization performance has a PR curve substantially above this constant, at most levels of recall.** The average precision (AP) of each detector is presented in table IV, **which also shows the number of features used by the discriminant saliency detector for each class.**

Three conclusions are supported by these results. First, all top-down pruning strategies improve the localization accuracy of bottom-up interest points, especially at low recall rates, indicating

Fig. 6. Saliency maps produced by DSD for various images and objects of interest. Bright pixels flag the most salient image locations.

that top-down pruning concentrates interest points on regions informative of object presence. **However, for all methods other than DSD-BU, precision drops quickly and is nearly constant at most levels of recall. This suggests that these detectors respond equally to object of interest and background. On the other hand, DSD-BU has much higher precision at most levels of recall, with a drop in performance only at very high recall levels. This indicates that DSD-BU is much more likely to produce interest points co-located with the object of interest. This hypothesis is confirmed by Figure 6, where we present saliency maps produced by DSD.** All scenes contain instances from two object classes, e.g. (a) “person” and “car”, (b) “person” and “bus”, (c) “motorbike” and ”car”, and (d) “person” and “motorbike”. **In each case, the groundtruth bounding box is shown (for reference) around the object of interest, on the left, while the saliency map for the detection of that object is shown on the right.** It is clear that DSD successfully switches between the two objects, highlighting the one of interest and suppressing all others.

Second, among the four top-down detectors, DSD-BU clearly achieves the best localization

Fig. 7. Salient locations at 40% recall. Top to bottom: original images (objects marked by their **groundtruth** bounding boxes), salient locations pruned by DSD, DVW, LSVM, and pLSA. Each circle represents the location and size of a salient point. White (black) indicates that the salient point falls inside (outside) the ground truth bounding box for the object.

accuracy. Its median AP, across the eight object classes, is at least 15% higher than those of DVW, LSVM, and pLSA. This suggests that all stages of a top-down saliency detector must be discriminant. The main difference is that, while DSD starts by selecting discriminant features, DVW, LSVM, and pLSA cluster interest points. Since neither interest points nor clustering is discriminant, much of the information of interest for recognition is eliminated before any discriminant learning takes place. While visual words are complex features, this non-discriminant nature limits their localization performance. In fact, Figure 5 shows that they can perform substantially worse than a simple orthogonal band-pass feature decomposition. Discriminant selection from a

Fig. 8. Salient locations at 40% recall. Top to bottom: original images (objects marked by their **groundtruth** bounding boxes), salient locations pruned by DSD, DVW, LSVM, and pLSA. Each circle represents the location and size of a salient point. White (black) indicates that the salient point falls inside (outside) the ground truth bounding box for the object.

poorly discriminant feature set does not guarantee good classification performance.

Comparing DVW and pLSA, DVW achieves a slightly better average performance. Again, this suggests an advantage for explicitly discriminant representations. Although pLSA is trained in a "discriminant" manner, it is intrinsically a generative model with limited ability to distinguish objects of interest from clutter. The observation that DVW also performs somewhat better than LSVM indicates an advantage of soft (the probabilistic representation of DVW) over hard (the visual word histogram of LSVM) assignments of interest points to visual words. Figures 7 and 8 present saliency detection examples for the four top-down detectors. **To produce these pictures,**

TABLE V

IMAGE CLASSIFICATION ACCURACY (%) OF BFSP CLASSIFIERS FOR PASCAL2006.

	bicycle	bus	car	cat	dog	horse	bike	person	median
Zhang [13]	94.8	98.1	97.5	93.7	87.6	92.6	96.9	85.5	94.8
BU	94.01	97.96	97.16	92.58	85.57	90.03	96.26	82.22	93.31
BU-GT	96.18	99.34	97.89	94.32	85.78	92.59	98.56	87.42	95.25
Gain (%)	2.17	1.38	0.73	1.74	0.21	2.56	2.3	5.2	1.96
DSD-BU	94.2	98.12	97.29	92.77	86.01	91.02	97.09	82.41	93.49
Norm. Gain (%)	8.8	11.6	17.8	10.9	209	38.7	36.1	3.7	14.7

the saliency map of each image was thresholded, and all bottom-up interest points of sub-threshold saliency was eliminated. In all cases, the threshold was set at the level of 40% **recall. The top image in each column of Figures 7, 8 presents an object and its ground-truth bounding box. The remaining images of the column display the interest points whose center falls within this box in white, and the rest in black.** Clearly, points pruned with DSD are more likely to be co-located with the target than those pruned with the other methods.

Finally, we note that although strong contextual influences exist for some classes, e.g. car and bus, DSD does not appear to have any problems learning features informative to the object of interest, rather than the background. This indicates that the consistency of appearance among the objects in these classes is higher than that of the background, making object features more discriminant than background features. This leads to their selection by discriminant saliency.

C. Image classification

We have also evaluated the benefits of DSD for image classification. These experiments were based on the BFSP classifier described in Section V-A.4. The classification accuracy obtained with the original interest points was compared to that achieved when these points were pruned by top-down saliency. Table V presents the AUC achieved for the different object classes. Three classification results are listed for each object category. They were obtained by varying the set of interest points fed to both K-means clustering and classifier. The first, labeled "BU", is the ensemble of all points generated by the interest point operators (HarrLap, HesLap, and DoG). This was the set of points adopted in the original implementation [13]. The second is the set of DSD-BU points, produced by pruning these interest points with discriminant saliency. To emphasize the areas of the objects of interest, saliency maps were first thresholded, by setting to

TABLE VI

COMPUTATION TIME, PER IMAGE, FOR EACH STAGE OF BU AND DSD-BU.

	Saliency detection	Quantization	SVM classifier	Overall
BU	$0s$	$9.59s$	$0.67s$	$10.3s$
DSD-BU	$0.197s$	$6.79s$	$0.67s$	$7.66s$

zero all locations of saliency smaller than 10% of the maximum (measured over the entire map). The third is the set of interest points that fall inside the ground truth bounding box (BU-GT). The three sets are derived from the same pool of interest points, but result from different pruning strategies: (1) no pruning, (2) discriminant and practical pruning, and (3) perfect but unrealistic pruning. **For completeness, we also include the results of [13]. Despite a significant effort on our part to replicate [13], our implementation (BU) did not match these results. Note, nevertheless, that BU-GT and DSD-BU are identical to BU up to pruning of interest points, making the comparison to BU the most relevant.**

As expected, the performance of BU-GT is usually the best. The only exception is the dog class where DSD-BU actually outperforms BU-GT. In all cases, BU performs the worst. To quantify the improvements of DSD-BU over BU, we normalized the gain from BU to DSD-BU by that from BU to BU-GT, i.e. $\frac{accuracy(\mathrm{DSD-BU})-accuracy(\mathrm{BU})}{accuracy(\mathrm{BU-GT})-accuracy(\mathrm{BU})}$. This measure reflects the fact that BU-GT is expected to achieve the best performance (since it is based on perfect object segmentation) and quantifies the percentage of the gap between BU and BU-GT which is recovered by DSD-BU. As can be seen from the last row of Table V, the median value of this percentage gain across classes is 14.7. For classes of dog, motorbike, and horse, DSD-BU recovers more than 36% of the gap between BU and BU-GT. These results show that the proposed discriminant saliency detector captures relevant information for object classification. Finally, it is worth pointing out that, for all object classes, DSD pruned away at least 30% of the BU points. **As shown in table VI, this translates into non-negligible computational savings for image classification, where the processing time is dominated by the search for the visual word closest to the SIFT descriptor extracted at each interest point (an operation that we refer to as *quantization*). As can be seen from the table, the savings in classification time are proportional to the percentage of interest points pruned by top-down saliency ($\approx 30\%$).** This can be of interest for applications where computation is limited.

Fig. 9. Saliency maps (bottom row) obtained on various textures (shown in top row) from Brodatz. Bright pixels flag salient locations.

D. Diversity of salient visual attributes

We finalize with a qualitative experiment, based on the Brodatz texture database [63], which illustrates the diversity of visual attributes that can be declared salient by discriminant saliency. Brodatz contains 112 texture classes, each represented by nine images. These classes include a great variety of salient attributes (e.g. corners, contours, regular geometric figures (circles, squares, etc.), texture gradients, crisp and soft edges). The database was divided into a training and a test set, using a set-up commonly adopted for texture retrieval (described in detail in [64]). The salient features of each class were computed from the training set, and the test images used to produce all saliency maps. The process was repeated for all texture classes, on a one-vs-all setting with each class sequentially considered as the "one" class. As illustrated by Figure 9, discriminant saliency has no difficulty in 1) ignoring highly textured background areas in favor of a more salient foreground object (two leftmost images), which could itself be another texture, 2) detecting as salient a wide variety of shapes, contours of different crispness and scale, or 3) even assigning strong saliency to texture gradients (rightmost image). This robustness is a consequence of the fact that salient features are selected according to both the class of interest and the set of images in the *all* class.

VI. Conclusion and future work

We have proposed a novel formulation for top-down saliency, denoted as discriminant saliency, which is intrinsically grounded on the recognition problem. Under this formulation, salient features are those that best discriminate between visual stimuli drawn from a target class and those drawn from a null hypothesis, composed of all other classes of possible recognition interest. Saliency is then defined as the confidence with which locations in the visual field can be classified as containing stimuli drawn from the target class. While both discrimination and classification confidence can be defined with respect to a number of previously proposed computational principles for perceptual organization, we have argued for the adoption of Barlow's principle of inference by detection of suspicious coincidences. In particular, it was shown that the combination of this principle with known statistical properties of natural images enables computationally parsimonious implementations of both feature selection and saliency detection. The resulting discriminant saliency detector is quite effective as a focus-of-attention mechanism, which can prune the interest points commonly used in computer vision according to their relevance for recognition. Experiments were designed to evaluate the benefits of this focus-of-attention mechanism in object localization and image classification tasks. In both cases, the addition of discriminant saliency was shown to improve the performance of current state-of-the-art methods.

Regarding future work, a number of questions merit further exploration. First, it would be interesting to augment the current saliency detector with a more diverse set of candidate features. For example, for objects with characteristic shapes, such as sheep or cows, the addition of contour-based features (such as adjacent contour segments [65]) should prove beneficial. On the other hand, the feature pool can be enriched by relying on more sophisticated forms of feature extraction, such as the complex template features of [42]. Some preliminary ideas in this direction have been studied in [66], [67]. The goal is to include feature learning within the model-building loop, so as to directly optimize the overall classification accuracy. The overarching point is that saliency should always be driven by *optimality with respect to the end task*. Second, it would be interesting to design object recognition systems that rely on discriminant saliency to *precisely* segment the object of interest from surrounding clutter. While we have shown that discriminant saliency is able to reliably locate objects of interest, it currently does not produce a finely tuned segmentation mask. This will require the combination of top-down saliency with bottom-up *image segmentation* algorithms, and is a topic for future research.

Acknowledgments

The authors would like to acknowledge Dr. Jianguo Zhang for suggestions on the implementation of the classifier of [13]. The work was partially funded by NSF award IIS-0448609.

REFERENCES

[1] A. Yarbus, *Eye movements and vision*. New York: Plenum, 1967.

[2] S. E. Palmer, *Vision Science: Photons to Phenomenology*. The MIT Press, 1999.

[3] C. Koch and S. Ullman, "Shifts in selective visual attention: towards the underlying neural circuitry," *Hum. Neurobiol.*, vol. 4, pp. 219–227, 1985.

[4] J. M. Wolfe, "Guided search 2.0: A revised model of visual search," *Psychonomic Bulletin & Review*, vol. 1, no. 2, pp. 202–238, 1994.

[5] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Trans. PAMI*, vol. 20, no. 11, 1998.

[6] J. K. Tsotsos, S. M. Culhane, W. Y. K. Winky, Y. Lai, N. Davis, and F. Nuflo, "Modeling visual attention via selective tuning," *Artif. Intell.*, vol. 78, no. 1-2, pp. 507–545, 1995.

[7] L. Itti, "Automatic foveation for video compression using a neurobiological model of visual attention," *IEEE Trans. on Image Processing*, vol. 13, pp. 1304–1318, 2004.

[8] D. Walther and C. Koch, "Modeling attention to salient proto-objects," *Neural Networks*, vol. 19, pp. 1395–1407, 2006.

[9] F. Shic and B. Scassellati, "A behavioral analysis of computational models of visual attention," *International Journal of Computer Vision*, vol. 73, pp. 159–177, 2007.

[10] R. Fergus, P. Perona, and A. Zisserman, "Object class recognition by unsupervised scale-invariant learning," in *Proc. IEEE Conf. CVPR*, 2003.

[11] J. Sivic, B. C. Russell, A. A. Efros, A. Zisserman, and W. T. Freeman, "Discovering objects and their localization in images," in *Proc. IEEE ICCV*, 2005, pp. 370–377.

[12] R. Fergus, L. Fei-Fei, P. Perona, and A. Zisserman, "Learning object categories from google's image search," in *Proc. IEEE ICCV*, 2005, pp. 1816–1823.

[13] J. Zhang, M. Marszalek, S. Lazebnik, and C. Schmid, "Local features and kernels for classification of texture andobject categories: A comprehensive study," *International Journal of Computer Vision*, vol. 73, no. 2, pp. 213–238, 2007.

[14] O. Chum and A. Zisserman, "An exemplar model for learning object classes," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, vol. 1, 2007, pp. 1–8.

[15] P. Quelhas, F. Monay, J.-M. Odobez, D. Gatica-Perez, and T. Tuytelaars, "A thousand words in a scene," *IEEE Trans. on PAMI*, vol. 29, no. 9, pp. 1575–1589, 2007.

[16] G. Dorkó and C. Schmid, "Selection of scale-invariant parts for object class recognition," in *International Conference on Computer Vision*, vol. 1, 2003, pp. 634–640.

[17] A. Opelt, A. Pinz, M. Fussenegger, and P. Auer, "Generic object recognition with boosting," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 28, no. 3, pp. 416–431, March 2006.

[18] C. Harris and M. Stephens, “A combined corner and edge detector,” in *Alvey Vision Conference*, 1988, pp. 147–151.

[19] W. Förstner, “A framework for low level feature ex-traction,” in *Proc. of ECCV*, 1994, pp. 383–394.

[20] A. Sha’ashua and S. Ullman, “Structural saliency: the detection of globally salient structures using a locally connected network,” in *Proc. ICCV*, 1988, pp. 321–327.

[21] H. Asada and M. Brady, “The curvature primal sketch,” *IEEE Trans. PAMI*, vol. 8, no. 1, pp. 2–14, 1986.

[22] D. Reisfeld, H. Wolfson, and Y. Yeshurun, “Context-free attentional operators: The generalized symmetry transform,” *Intl J. Comp. Vis.*, vol. 14, pp. 119–130, 1995.

[23] G. Heidemann, “Focus-of-attention from local color symmetries,” *IEEE Trans. PAMI*, vol. 26, no. 7, pp. 817–830, 2004.

[24] D. G. Lowe, “Object recognition from local scale-invariant features,” in *Proc. ICCV*, 1999, pp. 1150–1157.

[25] T. Lindeberg, “Scale-space theory: A basic tool for analyzing structures at different scales,” *J. Applied Statistics*, vol. 21, no. 2, pp. 224–270, 1994.

[26] K. Mikolajczyk and C. Schmid, “Scale and affine invariant interest point detectors,” *Int’l J. Comp. Vis.*, vol. 60, no. 1, pp. 63–86, 2004.

[27] T. Kadir, A. Zisserman, and M. Brady, “An affine invariant saliency region detector,” in *Proc. ECCV*, 2004, pp. 228–241.

[28] K. Yamada and G. W. Cottrell, “A model of scan paths applied to face recognition,” in *Proceedings of the Seventeenth Annual Cognitive Science Conference*. Pittsburgh, PA: Mahwah: Lawrence Erlbaum, 1995, pp. 55–60.

[29] N. Sebe and M. S. Lew, “Comparing salient point detectors,” *Pattern Recognition Letters*, vol. 24, no. 1-3, pp. 89–96, Jan. 2003.

[30] T. Kadir and M. Brady, “Scale, saliency and image description,” *International Journal of Computer Vision*, vol. 45, pp. 83–105, Nov. 2001.

[31] S. Lazebnik, C. Schmid, and J. Ponce, “Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories,” in *Proc. IEEE CVPR*, 2006.

[32] E. Nowak, F. Jurie, and B. Triggs, “Sampling strategies for bag-of-features image classification,” in *9th European Conference on Computer Vision*. Springer, 2006, pp. 490–503.

[33] R. Linsker, “Self-organization in a perceptual network,” *IEEE Computer*, vol. 21, no. 3, pp. 105–117, 1988.

[34] F. Attneave, “Informational Aspects of Visual Perception,” *Psychological Review*, vol. 61, pp. 183–193, 1954.

[35] H. Barlow, “Redundancy Reduction Revisited,” *Network: Computation in Neural Sysytems*, vol. 12,

pp. 241–253, 2001.

[36] H. B. Barlow, "Cerebral cortex as a model builder," in *Models of the Visual Cortex*, V. D. D. Rose, Ed. New York: John Wiley, 1985, p. 37C46.

[37] D. Gao and N. Vasconcelos, "Decision-theoretic saliency: computational principle, biological plausibility, and implications for neurophysiology and psychophysics," *Neural Computation*, vol. 21, pp. 239–271, 2009.

[38] D. Gao, V. Mahadevan, and N. Vasconcelos, "On the plausibility of the discriminant center-surround hypothesis for visual saliency," *Journal of Vision*, vol. 8, no. 7, pp. 1–18, 6 2008. [Online]. Available: http://journalofvision.org/8/7/13/

[39] R. Duda, P. Hart, and D. Stork, *Pattern Classification*. John Wiley & Sons, 2001.

[40] R. Battiti, "Using Mutual Information for Selecting Features in Supervised Neural Net Learning," *IEEE Trans. Neural Networks*, vol. 5, no. 4, pp. 537–550, July 1994.

[41] N. Vasconcelos, "Feature selection by maximum marginal diversity," in *Proc. NIPS*, Vancouver, Canada, 2002.

[42] M. Vidal-Naquet and S. Ullman, "Object recognition with informative features and linear classification," in *Proc. ICCV*, Nice, France, 2003.

[43] M. Vasconcelos and N. Vasconcelos, "Natural image statistics and low complexity feature selection," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, In press.

[44] T. O. Binford, "Inferring surfaces from images," *Artificial Intelligence*, vol. 17, no. 1-3, pp. 205–244, 1981.

[45] D. G. Lowe, "Three-dimensional object recognition from single two-dimensional images," *Artificial Intelligence*, vol. 31, no. 3, pp. 355–395, 1987.

[46] R. Buccigrossi and E. Simoncelli, "Image compression via joint statistical characterization in the wavelet domain," *IEEE Trans. on Image Processing*, vol. 8, pp. 1688–1701, 1999.

[47] J. Huang and D. Mumford, "Statistics of Natural Images and Models," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 1999, pp. 541–547.

[48] A. Treisman and G. Gelade, "A feature-integration theory of attention," *Cognitive Psychology*, vol. 12, no. 1, pp. 97–136, 1980.

[49] M. I. Posner, "Orientation of attention," *Quart. J. Experimental Psychology*, vol. 32, pp. 3–25, 1980.

[50] J. W. Modestino, "Adaptive nonparametric detection techniques," in *Nonparametric Methods in Communications*, P. Papantoni-Kazakos and D. Kazakos, Eds. New York: Marcel Dekker, 1977, pp. 29–65.

[51] S. G. Mallat, "A theory for multiresolution signal decomposition: The wavelet representation," *IEEE Trans. PAMI*, vol. 11, no. 7, pp. 674–693, 1989.

[52] K. Sharifi and A. Leon-Garcia, "Estimation of shape parameter for generalized gaussian distributions

in subband decompositions of video," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 5, no. 1, pp. 52–56, 1995.

[53] M. N. Do and M. Vetterli, "Wavelet-based texture retrieval using generalized gaussian density and kullback-leibler distance," *IEEE Trans. IP*, vol. 11, pp. 146–158, 2002.

[54] M. Everingham, A. Zisserman, C. K. I. Williams, and L. Van Gool, "The PASCAL Visual Object Classes Challenge 2006 (VOC2006) Results," http://www.pascal-network.org/challenges/VOC/voc2006/results.pdf.

[55] D. Gao and N. Vasconcelos, "An experimental comparison of three guiding principles for the detection of salient image locations: Stability, complexity, and discrimination," in *The 3rd International Workshop on Attention and Performance in Computational Vision*, L. Paletta and E. Rome, Eds. Berlin Heidelberg: Springer-Verlag, 2007, pp. 184–197.

[56] A. Torralba, "Contextual priming for object detection," *International Journal of Computer Vision*, vol. 53, no. 2, pp. 169–191, 2003.

[57] A. Oliva and A. Torralba, "The role of context in object recognition," *Trends in Cognitive Sciences*, vol. 11, no. 12, pp. 520–527, 2007.

[58] M. Vasconcelos, N. Vasconcelos, and G. Carneiro, "Weakly supervised top-down image segmentation," in *Proc. of IEEE CVPR*, 2006, pp. 1001–1006.

[59] C. Bouveyron, J. Kannala, C. Schmid, and S. Girard, "Object localization by subspace clustering of local descriptors," in *ICVGIP*, 2006.

[60] F. Jurie and B. Triggs, "Creating efficient codebooks for visual recognition," in *International Conference on Computer Vision (ICCV)*, 2005.

[61] D. Mladenić, J. Brank, M. Grobelnik, and N. Milic-Frayling, "Feature selection using linear classifier weights: interaction with classification models," in *Proc. ACM SIGIR conf. on Research and development in information retrieval*, 2004, pp. 234–241.

[62] T. Hofmann, "Unsupervised learning by probabilistic latent semantic analysis," *Mach. Learn.*, vol. 42, pp. 177–196, 2001.

[63] P. Brodatz, *Textures: A Photographic Album for Artists and Designers*. Dover, NewYork, 1966.

[64] N. Vasconcelos and G. Carneiro, "What is the role of independence for visual regognition?" in *Proc. ECCV*, Copenhagen, Denmark, 2002.

[65] V. Ferrari, L. Fevrier, F. Jurie, and C. Schmid, "Groups of adjacent contour segments for object detection," *IEEE Trans. on PAMI*, vol. 30, no. 1, pp. 36–51, 2008.

[66] D. Gao and N. Vasconcelos, "Integrated learning of saliency, complex features, and object detectors from cluttered scenes," in *Proc. IEEE CVPR*, 2005, pp. 282–287.

[67] S. Han and N. Vasconcelos, "Complex discriminant features for object classification," in *Proc. IEEE International Conference on Image Processing (ICIP)*, San Diego, USA, 2008.