A Bayesian Architecture for Combining Saliency Detectors

Dashan Gao and Nuno Vasconcelos

Statistical Visual Computing Laboratory

SVCL UCSD

SVCL-TR 2005/01

June 2005

A Bayesian Architecture for Combining Saliency Detectors

Dashan Gao and Nuno Vasconcelos
Statistical Visual Computing Lab
Department of Electrical and Computer Engineering
University of California, San Diego

June 2005

Abstract

Saliency mechanisms can play an important role in the ability of recognition systems to deal with cluttered scenes. Saliency detection has also been an active area of computer vision, where existing solutions can be divided into two major classes: domain-independent and domain-dependent. In this work, it is proposed that the two classes are complementary and can be addressed within a unified formulation of the saliency problem, inspired on regularization theory. A Bayesian saliency framework is then proposed, in which domain-independent saliency maps are interpreted as priors for salient location, which can be used to regularize estimates of salient point location derived with domain-dependent procedures. Saliency maps are modeled as mixture distributions, and an analytical solution derived for the posterior distribution of true salient locations given the observed saliency measurements. This framework is shown to enable explicit control over the relative importance of the two types of saliency, reveals an interpretation of domain-dependent saliency as a focus-of-attention mechanism, and has a simple non-parametric extension. Experimental evaluation demonstrates the benefits of Bayesian saliency for weakly supervised recognition problems.

Author email: dgao@ucsd.edu

SVCL Technical reports are available on the SVCL's web page at
http://www.svcl.ucsd.edu

University of California, San Diego
Statistical Visual Computing Laboratory
9500 Gilman Drive, Mail code 0407
EBU 1, Room 5512
La Jolla, CA 92093-0407

1 Introduction

The formulation of recognition as a problem of statistical classification has enabled significant progress in the area, over the last decades. Recently, there has been growing interest in the problem of recognition from cluttered examples [1–3, 13, 19, 20]. While increasing the complexity of visual recognition, the ability to recognize objects in the presence of clutter opens the possibility for training recognizers with weak supervision. For example, instead of requiring a set of training faces cropped into 20×20 pixel arrays, with all hair precisely cropped out, lighting gradients removed, and so on, a weakly supervised face detector would simply be trained from a set of images containing faces. Since preparation of examples is the bulk of the effort currently required to train any recognizer, the potential practical impact of weakly supervised recognition is substantial.

The ability to learn in the presence of substantial amounts of clutter is also a hallmark of biological vision. One property that plays an important role in this ability is the existence of saliency mechanisms: biological vision systems rarely have to exhaustively scan a scene in order to detect an object of interest. Instead, salient locations simply pop-out in result of the operation of pre-recognition saliency mechanisms. Replicating this ability to find salient points has been a goal of computer vision for some decades.

Broadly speaking, existing saliency detectors can be divided into two major classes: *domain-independent* and *domain-dependent*. The most popular formulations in the first class are probably those that treat saliency as the *detection of specific visual attributes*, such as edges or corners (also called "interest points"). Popular examples of detectors in this class are those by Harris [4] and Föstner [5]. Other possible saliency definitions are based on 1) more generic and *data-driven* saliency measures, such as image complexity [6–8], or 2) *models of biological vision* [9, 10]. The unifying theme for all these formulations is a definition of saliency in terms of properties which, while universally desirable for a saliency detector, do not take into account the target application-domain, e.g. recognition. Instead, they pose saliency as an end in itself.

Domain-dependent methods ground the definition of saliency in *specific application-domains*. This class includes, for example, recent proposals to ground saliency on the recognition problem, by equating *saliency to discriminant power* [3, 11–13]. Under this formulation, 1) salient locations are those that best differentiate the visual class of interest from all others, and 2) saliency relies on a preliminary stage of feature selection, based on how discriminant each feature is with respect to the visual class to recognize.

Both principles have their advantages and limitations. Domain-independent detectors can be made optimal with respect to universally desirable properties for saliency detection. For example, Harris and Föstner produce salient points that are optimally stable with respect to geometric image transformations. They also do not require training and have low-complexity. On the other hand, the absence of an application-domain restricts the detectors in this class to being optimal in very generic senses, and the resulting interest points are rarely the best for specific applications. In the Harris/Föstner example they are simply corners.

With respect to biological vision, human experiments have shown that, even for relatively straightforward saliency tasks, where subjects are 1) shown images that they have already seen and 2) simply asked to point out salient regions, different people do not seem to agree on more than about 50% of the salient locations [10]. This seems to rule out the exclusive adoption, by biological vision systems, of universal (domain-independent) saliency principles, even when no domain-specific goals are explicitly set for the saliency task. Recent results in discriminant saliency detection [13] have shown that domain-dependent saliency can be beneficial for computer vision as well. For example, it appears clear that, in the recognition domain, discriminant saliency produces salient points that are substantially more informative than those produced by domain-independent methods.

While these observations indicate that some form of domain-specificity is always likely to be beneficial, there is inconvenience in the adoption of domain-dependent principles. In particular, because such principles are by definition based on learning, their performance depends on the size of the available training set. For small training sets, domain-specificity can lead to over-fitting and poor saliency detection. In this case, the properties that domain-independent principles enforce, e.g. stability, become desirable, as a way to *regularize* the domain-dependent solution. Regularization is widely used in statistical learning, as a means to improve generalization guarantees: a regularized learner requires a much smaller number of training examples to achieve the performance of an equivalent learner without regularization [21]. Under the regularization point of view, domain-dependent and -independent saliency principles are complementary, rather than competitors.

In this work, we propose a *Bayesian regularization framework* for the *fusion of saliency maps of the two types*. We regard the output of a domain-independent saliency detector as a prior for salient locations, and the output of a domain-dependent saliency detector as a set of saliency observations. The location uncertainty associated with each salient point is encoded as a Gaussian distribution over image coordinates, leading to an approximation of each saliency map by a mixture distribution. This enables the derivation of an analytical solution for the posterior distribution of the true, but unknown, salient locations. This posterior is shown to be another Gaussian mixture, and is completely characterized by the derivation of its parameters from those associated with the two saliency maps.

The posterior distribution for the true salient locations is also shown to have various interesting properties. First, it consists of an intuitive combination of all terms in the two saliency distributions. Second, it enables the introduction of a regularization constant that allows explicit control of the relative importance of each distribution in the posterior saliency estimates. Third, it suggests the interpretation of domain-dependent saliency as a focus-of-attention mechanism which suppresses domain-independent salient points that are not informative for the domain of interest. Finally, it has a non-Bayesian interpretation as the simple multiplication of the two saliency maps, which enables a non-parametric extension of trivial computational complexity.

The advantages of Bayesian saliency, over the application of either domain-independent or domain-dependent saliency in isolation, are illustrated for two challenging aspects of weakly supervised recognition: 1) its implementation in the presence of substantial amounts of clutter, and 2) the robustness of salient points for segmenting examples. the ability to automatically segment examples for training traditional object detectors.

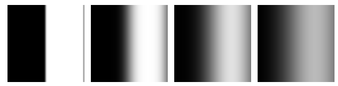


Figure 1: Edge location uncertainty vs. scale. From left to right: as scale increases (and resolution decreases) edge location is less certain.

In both cases, Bayesian saliency achieves the best performance, also outperforming state-of-the-art methods recently proposed in the literature.

2 Bayesian saliency

In this section we derive a Bayesian solution for the fusion of domain-dependent ("DD") and domain-independent ("DI") saliency maps. We start by the case where both maps have a single salient point, then consider the case of multiple DI salient points and a single DD salient point, and finally address the general case where both maps have multiple salient points.

2.1 Single salient point

A salient point $\mathbf{s}$ is characterized by three parameters: its saliency strength α, image location $\mathbf{x}$, and scale σ. In this work, it is assumed that both the strength and scale are known[1]. If the application, to the image, of a DD saliency detector results in a salient point of scale σ^{dd}, the observed salient location is modeled as a Gaussian random variable $\mathbf{X} = (x, y)$ of covariance $\mathbf{\Sigma} = (\sigma^{dd})^2\mathbf{I}$ and centered on the true, but unknown, salient location μ,

$$P_{\mathbf{X}|\mu}(\mathbf{x}|\mu) = \mathcal{G}(\mathbf{x}, \mu, (\sigma^{dd})^2\mathbf{I}).$$

where $\mathcal{G}(\mathbf{x}, \mu, \sigma^2\mathbf{I}) = \frac{1}{2\pi\sigma^2} \exp\left(-\frac{(\mathbf{x}-\mu)^T(\mathbf{x}-\mu)}{2\sigma^2}\right)$, and $\mathbf{I}$ is the 2×2 identity matrix. This reflects the fact that the location uncertainty is larger for salient points at larger image scales (low resolution) than for points at small scales (high resolution) , as illustrated in Figure 1[2].

As is usual in Bayesian inference, the uncertainty about the true location μ is formalized by considering this parameter a random variable and introducing a prior distribution $P_\mu(\mu)$. As discussed above, it is sensible to derive this prior from a DI saliency principle. Assuming that a DI saliency detector produced a salient point

[1] While, in practice, this is not strictly true, there is usually a fair amount of tolerance to errors in these parameters. For example, it is common to simply classify points as salient or non-salient, in which case a measure of saliency strength is not even required. With respect to the scale parameter, it is common practice to consider only a finite set of possible scales. Since the selection of the best among these with small error is usually feasible, the assumption of known scale is a reasonable one. In future work we will consider a fully Bayesian solution that takes these parameters as random variables as well.

[2] More complex models are obviously possible, e.g. a covariance structure that assigns more uncertainty along, than across, edges. The extension of the framework now proposed so as to support them is straightforward.

$\mathbf{s}^{di} = (\alpha^{di}, \mu^{di}, \sigma^{di})$, the prior density for location is, once again, assumed Gaussian

$$P_\mu(\mu) = \mathcal{G}(\mu, \mu^{di}, (\sigma^{di})^2\mathbf{I}).$$

The location $\mathbf{x}^{dd}$ of the DD salient point is then viewed as an observation of $\mathbf{X}$, leading to a posterior distribution for the true salient location of the form [17]

$$P_{\mu|\mathbf{X}}(\mu|\mathbf{x}^{dd}) = \mathcal{G}(\mu, \mu^s, (\sigma^s)^2\mathbf{I}), \tag{1}$$

with

$$\mu^s = \frac{(\sigma^{di})^2}{(\sigma^{di})^2 + (\sigma^{dd})^2}\mathbf{x}^{dd} + \frac{(\sigma^{dd})^2}{(\sigma^{di})^2 + (\sigma^{dd})^2}\mu^{di}, \tag{2}$$

$$(\sigma^s)^2 = \frac{(\sigma^{di})^2(\sigma^{dd})^2}{(\sigma^{di})^2 + (\sigma^{dd})^2}. \tag{3}$$

The relative importance of the DD and DI saliency maps, can be controlled by multiplying the prior variance by a regularization constant σ, i.e. by replacing σ^{di} with $\sigma \cdot \sigma^{di}$ in the equations above. Note that, as $\sigma \to \infty$, $\mu^s = \mathbf{x}^{dd}$ and $\sigma^s \to \sigma^{dd}$, making the posterior distribution equal to the Gaussian associated with the DD salient point $\mathbf{s}^{dd}$. On the other hand, when $\sigma \to 0$, $\mu^s = \mu^{di}$ and $\sigma^s \to 0$, making the posterior distribution equal to the delta function centered in the location of the DI salient point μ^{di}. This is illustrated by Figure 2 where we combine the most salient point produced by a (DI) Harris detector with the most salient point produced by the (DD) discriminant saliency detector of [13].

2.2 Multiple domain-independent salient points

If there are various DI salient points $\{\mathbf{s}_1^{di}, \ldots, \mathbf{s}_n^{di}\}$, any of them could be responsible for the observed salient location $\mathbf{x}^{dd}$ produced by the DD saliency detector. To account for this we introduce a hidden variable Y, such that $Y = k$ when $\mathbf{s}_k^{di}$ is the responsible DI salient point, and the following generative model:

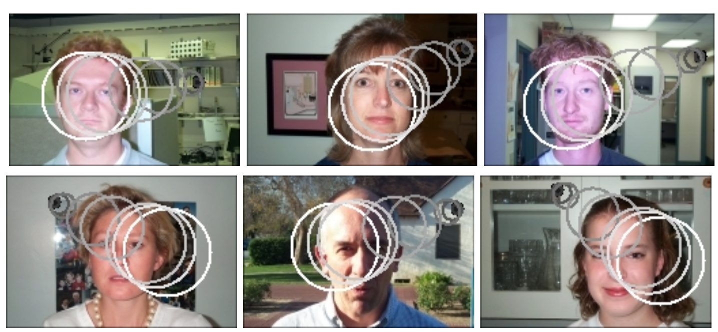


Figure 2: The posterior distribution (circle) of the most salient location as a function of the regularization constant σ. Brighter circles indicate larger values of σ: in all images the black (white) circle represents the most salient point detected by the domain-independent (dependent) detector.

1. the k^{th} DI salient point is chosen with probability $P_Y(k) = \frac{\alpha_k^{di}}{\sum_j \alpha_j^{di}}$.
2. the prior density for location becomes $P_{\mu|Y}(\mu|k) = \mathcal{G}(\mu, \mu_k^{di}, (\sigma_k^{di})^2\mathbf{I})$.
3. the observed salient location $\mathbf{x}^{dd}$ is sampled from the distribution $P_{\mathbf{X}|\mu}(\mathbf{x}|\mu)$.

The posterior distribution for the unknown salient location becomes

$$\begin{aligned}
P_{\mu|\mathbf{X}}(\mu|\mathbf{x}^{dd}) &= \frac{\sum_k P_{\mathbf{X},\mu,Y}(\mathbf{x}^{dd}, \mu, k)}{P_{\mathbf{X}}(\mathbf{x}^{dd})} \\
&= \frac{\sum_k P_{\mathbf{X}|\mu,Y}(\mathbf{x}^{dd}|\mu, k) P_{\mu|Y}(\mu|k) P_Y(k)}{P_{\mathbf{X}}(\mathbf{x}^{dd})} \\
&= \frac{\sum_k P_{\mathbf{X}|\mu}(\mathbf{x}^{dd}|\mu) P_{\mu|Y}(\mu|k) P_Y(k)}{\sum_j P_{\mathbf{X}|Y}(\mathbf{x}^{dd}|j) P_Y(j)}
\end{aligned}$$

where,

$$\begin{aligned}
P_{\mathbf{X}|\mu}(\mathbf{x}^{dd}|\mu) P_{\mu|Y}(\mu|k) &= \mathcal{G}(\mathbf{x}^{dd}, \mu, (\sigma^{dd})^2\mathbf{I}) \mathcal{G}(\mu, \mu_k^{di}, (\sigma_k^{di})^2\mathbf{I}) \\
&= \mathcal{G}(\mu, \mu_k^s, (\sigma_k^s)^2\mathbf{I}) \mathcal{G}(\mu_k^{di}, \mathbf{x}^{dd}, [(\sigma^{dd})^2 + (\sigma_k^{di})^2]\mathbf{I})
\end{aligned}$$

with

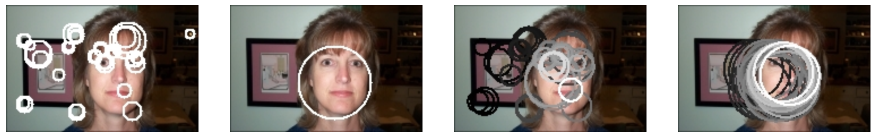
$$\mu_k^s = \frac{(\sigma_k^{di})^2}{(\sigma_k^{di})^2 + (\sigma^{dd})^2}\mathbf{x}^{dd} + \frac{(\sigma^{dd})^2}{(\sigma_k^{di})^2 + (\sigma^{dd})^2}\mu_k^{di} \tag{4}$$

$$(\sigma^s)^2 = \frac{(\sigma_k^{di})^2 (\sigma^{dd})^2}{(\sigma_k^{di})^2 + (\sigma^{dd})^2}. \tag{5}$$

and

$$\begin{aligned}
P_{\mathbf{X}|Y}(\mathbf{x}^{dd}|j) &= \int_\mu P_{\mathbf{X}|\mu,Y}(\mathbf{x}^{dd}|\mu, j) P_{\mu|Y}(\mu|j) \mathrm{d}\mu \\
&= \int_\mu P_{\mathbf{X}|\mu}(\mathbf{x}^{dd}|\mu) P_{\mu|Y}(\mu, j) \mathrm{d}\mu \\
&= \int_\mu \mathcal{G}(\mathbf{x}^{dd}, \mu, (\sigma^{dd})^2\mathbf{I}) \mathcal{G}(\mu, \mu_j^{di}, (\sigma_j^{di})^2\mathbf{I}) \mathrm{d}\mu \\
&= \mathcal{G}(\mu_j^{di}, \mathbf{x}^{dd}, [(\sigma^{dd})^2 + (\sigma_j^{di})^2]\mathbf{I}).
\end{aligned}$$

In this derivation, we have used known properties of the Gaussian distribution [17] and the fact that, given the true location, the observed location $\mathbf{x}^{dd}$ does not depend on the



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

Figure 3: Modulation of the focus of attention mechanism, associated with domain-dependent saliency, by σ. Images show salient locations detected by (a) Harris, (b) discriminant, (c) Bayesian ($\sigma^2 = 6$), and (d) Bayesian ($\sigma^2 = 200$) saliency detector. Brighter circles indicate stronger saliency.

prior, $P_{\mathbf{X}|\mu,Y}(\mathbf{x}^{dd}|\mu, k) = P_{\mathbf{X}|\mu}(\mathbf{x}^{dd}|\mu)$. It follows that

$$P_{\mu|\mathbf{X}}(\mu|\mathbf{x}^{dd}) = \sum_k \mathcal{G}(\mu, \mu_k^s, (\sigma_k^s)^2\mathbf{I})\pi(\mathbf{x}^{dd}, \mathbf{s}_k^{di}) \tag{6}$$

with

$$\pi(\mathbf{x}^{dd}, \mathbf{s}_k^{di}) = \frac{\mathcal{G}(\mu_k^{di}, \mathbf{x}^{dd}, [(\sigma^{dd})^2 + (\sigma_k^{di})^2]\mathbf{I})\alpha_k^{di}}{\sum_j \mathcal{G}(\mu_j^{di}, \mathbf{x}^{dd}, [(\sigma^{dd})^2 + (\sigma_j^{di})^2]\mathbf{I})\alpha_j^{di}},$$

and μ_k^s and σ_k^s as given in (2) and (3) with μ^{di} and σ^{di} replaced by μ_k^{di} and σ_k^{di} respectively.

It is interesting to compare this distribution to that of the case of a single DI salient point: the posterior is now a mixture of Gaussians of the form of (1), each weighted according to the link function $\pi(\mathbf{x}^{dd}, \cdot)$. Up to a constant, this is a Gaussian centered on the observed salient location $\mathbf{x}^{dd}$ produced by the DD detector, and penalizes the contributions of DI salient points which are located far from this observation. It enables the interpretation of the DD saliency detector as a *focus of attention* operator that suppresses DI salient points which are not discriminant for the object of interest.

As before, the relative importance of the DI and DD saliency maps can be controlled by multiplying all prior variances by a regularization constant σ, i.e. by replacing σ_k^{di} with $\sigma \cdot \sigma_k^{di}, \forall k$, in the equations above. This can be used to modulate the focus of attention mechanism, as illustrated in Figure 3, in which we present the top DD (obtained with the discriminant saliency detector of [13]) and the 40 top DI salient points (obtained with Harris) for one image, and the posterior distribution for the salient location obtained with two values of σ. Note that, as σ increases, attention is more narrowly focused on the salient points located inside the object of interest, in this case a face.

2.3 Multiple domain-dependent and domain-independent salient points

We have, so far, shown that a DD salient point can be interpreted as a focus-of-attention operator that, given 1) a collection of DI salient points $\{\mathbf{s}_1^{di}, \ldots, \mathbf{s}_n^{di}\}$ and 2) an observed DD salient location $\mathbf{x}^{dd}$, produces a Bayesian estimate of the true, but unknown, salient location $P_{\mu|\mathbf{X}}(\mu|\mathbf{x}^{dd})$ of the form of (6). The DD salient point $\mathbf{s}^{dd} = (\alpha^{dd}, \mathbf{x}^{dd}, \sigma^{dd})$ associated with $\mathbf{x}^{dd}$ can, therefore, be viewed as an *attentional*

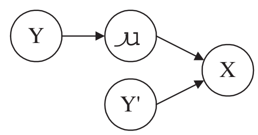


Figure 4: Graphical model for Bayesian saliency.

hypothesis about which image area is most likely to contain discriminant information for the object of interest.

Under this interpretation, a collection of DD salient points $\{\mathbf{s}_1^{dd}, \ldots, \mathbf{s}_m^{dd}\}$ is nothing more than a set of attentional hypotheses regarding the location of the target visual concept. This suggests the introduction of a second hidden variable Y', such that $Y' = l$ when the l^{th} attentional hypothesis holds, and the following generative model for salient locations:

1. the l^{th} attentional hypothesis is chosen with probability $P_{Y'}(l) = \frac{\alpha_l^{dd}}{\sum_j \alpha_j^{dd}}$.

2. a salient observation $\mathbf{x}_l^{dd}$ is then sampled according to the generative model in the previous section, conditioning all probabilities on the value of Y', i.e.,

$$P_{\mathbf{X}|\mu,Y'}(\mathbf{x}|\mu, l) = \mathcal{G}(\mathbf{x}, \mu, (\sigma_l^{dd})^2\mathbf{I}).$$

While it is conceivable that the attentional hypothesis would affect the prior density for location, e.g. by making some types of DI salient points more likely than others, we currently assume that this is not the case, i.e. $P_{\mu|Y,Y'}(\mu|k,l) = P_{\mu|Y}(\mu|k)$ and $P_{Y|Y'}(k|l) = P_Y(k)$. Hence, the second step of this procedure consists of the following sub-steps:

1. the k^{th} DI salient point is chosen with probability $P_Y(k) = \frac{\alpha_k^{di}}{\sum_j \alpha_j^{di}}$.

2. the prior density for location becomes $P_{\mu|Y}(\mu|k) = \mathcal{G}(\mu, \mu_k^{di}, (\sigma_k^{di})^2\mathbf{I})$.

3. the observed salient location $\mathbf{x}_l^{dd}$ is sampled from the distribution $P_{\mathbf{X}|\mu,Y'}(\mathbf{x}|\mu, l)$

$$P_{\mathbf{X}|\mu,Y'}(\mathbf{x}|\mu, l) = \mathcal{G}(\mathbf{x}, \mu, (\sigma_l^{dd})^2\mathbf{I}).$$

A graphical representation of this generative model is shown in Figure 4. where the conditional independence, $P_{\mathbf{X}|\mu,Y',Y}(\mathbf{x}_l^{dd}|\mu, l, k) = P_{\mathbf{X}|\mu,Y'}(\mathbf{x}_l^{dd}|\mu, l)$, can be read by the causal inference properties of Bayesian network [22]. It follows, from an analysis identical to that of the previous section, that the posterior for salient location, under the l^{th} attentional hypothesis, is

$$P_{\mu|Y',\mathbf{X}}(\mu|l, \mathbf{x}_l^{dd}) = \sum_k \mathcal{G}(\mu, \mu_{k,l}^s, (\sigma_{k,l}^s)^2\mathbf{I})\pi_l(\mathbf{x}_l^{dd}, \mathbf{s}_k^{di})$$

with

$$\mu^s_{k,l} = \frac{(\sigma^{di}_k)^2}{(\sigma^{di}_k)^2 + (\sigma^{dd}_l)^2}\mathbf{x}^{dd}_l + \frac{(\sigma^{dd}_l)^2}{(\sigma^{di}_k)^2 + (\sigma^{dd}_l)^2}\mu^{di}_k \tag{7}$$

$$(\sigma^s_{k,l})^2 = \frac{(\sigma^{di}_k)^2(\sigma^{dd}_l)^2}{(\sigma^{di}_k)^2 + (\sigma^{dd}_l)^2}. \tag{8}$$

$$\pi_l(\mathbf{x}, \mathbf{s}^{di}_k) = \frac{\mathcal{G}(\mu^{di}_k, \mathbf{x}, [(\sigma^{dd}_l)^2 + (\sigma^{di}_k)^2]\mathbf{I})\alpha^{di}_k}{\sum_j \mathcal{G}(\mu^{di}_j, \mathbf{x}, [(\sigma^{dd}_l)^2 + (\sigma^{di}_j)^2]\mathbf{I})\alpha^{di}_j}.$$

The overall posterior distribution is then

$$\begin{aligned} P_{\mu|\mathbf{X}}(\mu|\{\mathbf{x}^{dd}_1, \ldots, \mathbf{x}^{dd}_m\}) &= \frac{P_{\mathbf{X}|\mu}(\{\mathbf{x}^{dd}_1, \ldots, \mathbf{x}^{dd}_m\}|\mu)P_\mu(\mu)}{P_{\mathbf{X}}(\{\mathbf{x}^{dd}_1, \ldots, \mathbf{x}^{dd}_m\})} \\ &= \frac{\sum_l P_{\mathbf{X},Y'|\mu}(\mathbf{x}^{dd}_l, l|\mu)P_\mu(\mu)}{\sum_j P_{\mathbf{X},Y'}(\mathbf{x}^{dd}_j, j)} \qquad (9) \\ &= \sum_{k,l} \mathcal{G}(\mu, \mu^s_{k,l}, (\sigma^s_{k,l})^2\mathbf{I})\beta(\mathbf{x}^{dd}_l, \mathbf{s}^{di}_k) \qquad (10) \end{aligned}$$

with

$$\beta(\mathbf{x}^{dd}_l, \mathbf{s}^{di}_k) = \frac{\mathcal{G}(\mu^{di}_k, \mathbf{x}^{dd}_l, [(\sigma^{dd}_l)^2 + (\sigma^{di}_k)^2]\mathbf{I})\alpha^{di}_k\alpha^{dd}_l}{\sum_{i,j} \mathcal{G}(\mu^{di}_i, \mathbf{x}^{dd}_j, [(\sigma^{dd}_j)^2 + (\sigma^{di}_i)^2]\mathbf{I})\alpha^{di}_i\alpha^{dd}_j}.$$

where the omitted derivation from (9) to (10) is similar to that of section 2.2.

Note that this is a mixture of posterior distributions of the form of (6), i.e. a mixture of the $n \times m$ Gaussians associated with all pairs of DI and DD salient points. As before, the link function $\beta(\mathbf{x}^{dd}, \cdot)$ is, up to constants, a Gaussian centered on the observed salient location $\mathbf{x}^{dd}$ produced by the DD detector, and penalizes the contributions of DI salient points which are located far from it. The relative importance of the DD and DI saliency maps can still be controlled by multiplying all prior variances a regularization constant σ, i.e. replacing σ^{di}_k with $\sigma \cdot \sigma^{di}_k, \forall k$, in the equations above.

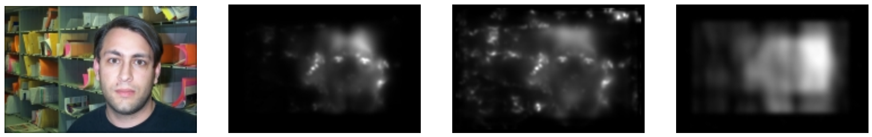
2.4 Non-parametric interpretation

An interesting low-level interpretation of the posterior distribution (10), that does not require Bayesian inference, can be obtained by noting that

$$\mathcal{G}(\mathbf{x}, \mu^{di}_k, (\sigma^{di}_k)^2\mathbf{I})\mathcal{G}(\mathbf{x}, \mathbf{x}^{dd}_l, (\sigma^{dd}_l)^2\mathbf{I}) = \mathcal{G}(\mathbf{x}, \mu^s_{k,l}, (\sigma^s_{k,l})^2\mathbf{I})\mathcal{G}(\mu^{di}_k, \mathbf{x}^{dd}_l, [(\sigma^{dd}_l)^2 + (\sigma^{di}_k)^2]\mathbf{I})$$

with $\mu^s_{k,l}$ and $(\sigma^s_{k,l})^2$ as given above. It follows that the posterior distribution of (10) is , up to constants, the product of the mixtures,

$$\sum_k \frac{\alpha^{di}_k}{\sum_i \alpha^{di}_i}\mathcal{G}(\mathbf{x}, \mu^{di}_k, (\sigma^{di}_k)^2\mathbf{I}), \quad \text{and} \quad \sum_l \frac{\alpha^{dd}_l}{\sum_i \alpha^{dd}_i}\mathcal{G}(\mathbf{x}, \mathbf{x}^{dd}_l, (\sigma^{dd}_l)^2\mathbf{I}),$$



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

Figure 5: Illustration of the non-parametric Bayesian saliency map (b) for the image of (a) derived from Harris (c) and discriminant saliency (d).

associated with the two saliency detectors, when the true salient locations are μ_k^{di} and $\mathbf{x}_l^{dd}$.

Noting that the mixture representation is a probabilistic approximation to the observed saliency maps, this enables a completely non-parametric representation of the posterior for the true salient location as the simple element-wise multiplication of the two saliency maps (plus normalization). This is illustrated by Figure 5. Note how, once again, the DD (discriminant) saliency map serves as a focus of attention mechanism to the more localized, but also more error-prone, DI (Harris) saliency map. What is lost, under this non-parametric interpretation, is the ability to introduce the regularization constant σ that modulates the strength of this focus-of-attention mechanism.

3 Segmentation of examples

The saliency maps of the previous section can be used to segment the regions associated with an object of interest from a collection of images. This, could be used to greatly expedite the process of training a traditional object detector (e.g., [15]). Rather than a collection of precisely segmented object views, the designer of the detector would simply provide a set of images where the objects appear, possibly surrounded by large amounts of clutter. Bayesian saliency could then be used to extract from this image set (and a set of images not containing the object) the examples required for training the traditional detector.

We next introduce a preliminary solution to this problem, based on simple template matching. We emphasize that the goal is not to fully solve the problem of segmenting training examples (a topic that we will address in future research) but to objectively evaluate whether the saliency maps resulting from Bayesian saliency provide robust and relevant information about the objects to be recognized. The example segmentation algorithm is as follows.

1. all images are preprocessed by homomorphic filtering for illumination normalization [16] and Gaussian low-pass filtering for noise reduction.

2. images from which an object is to be segmented are called *reference images*. For each reference image $\mathcal{R}$,

    (a) an image $\mathcal{M}$ is randomly selected, and its k salient locations sampled from the associated posterior Gauss mixture, as defined by (10). Each location

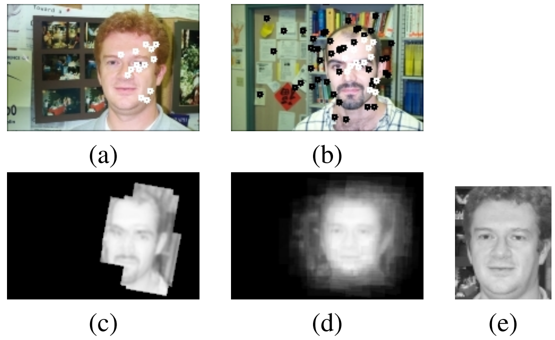


Figure 6: Illustration of the segmentation algorithm. (a) reference image; (b) matching image; (c) inliers mapped to the reference image plane; (d) average segmentation mask superimposed on image (a); (e) segmented patch. White and black circles in (a) and (b) represent, respectively, inliers and outliers returned by RANSAC.

inherits the scale of the Gaussian from which it was sampled. An image patch of that scale is cropped around the location.

(b) the peak correlation between each patch and $\mathcal{R}$ is computed. If it is greater than a threshold, T_{corr}, the patch is kept, otherwise discarded.

(c) an affine transformation is estimated from all matched patches using RANSAC [18]. If the number of the inliers returned by RANSAC is greater than a threshold, T_{in}, the transformation is kept. Otherwise, $\mathcal{M}$ is rejected, a rejection counter incremented, and step (2d) skipped.

(d) the inliers (patches) are mapped to the coordinate frame of $\mathcal{R}$, and a binary segmentation mask set to the union of the transformed regions of support.

(e) steps (2a) to (2d) are repeated until either a pre-set number of masks or rejections are produced.

(f) segmentation masks are averaged and the object is segmented from the locations of the reference image where the average mask is greater than a threshold, T_{mask}.

Figure 6 illustrates the segmentation process, marking rejected patches (outliers) with black circles and inliers with white circles. Inliers are, as shown in 6(c), mapped to the coordinate frame of the reference image (6(a)). Note that the segmentation mask produced by a single image (6(b)) can be poor due to the variations in appearance between the objects, or object views, depicted in the two images. Step (2e) guarantees a much more robust segmentation mask (shown in 6(d)). The segmented object patch is shown in 6(e).

4 Experiments

To evaluate the performance of Bayesian saliency we relied on the Caltech database, which has been proposed as a testbed for object detection in the presence of clutter [2].

Four image classes, faces (435 images of size 112x170), motorbikes (800 images of size 180x300), airplanes (800 images of size 200x300), and rear-cars (800 images of size 150x180), were used as the classes of interest, and a set of background images was also used as the negative class, with the experimental set up proposed in [2], .

4.1 Implementation of Bayesian saliency

Two representative saliency detectors, a (DI) Harris-Laplace (HarrLap) detector [14], and a (DD) discriminant saliency (DiscSal) detector [13], were selected to implement the Bayesian saliency detector [3]. The HarrLap detector searches points that are invariant to rotation and scale changes [14], and can be described as a sequence of two steps:

1. a scale-space representation with pre-selected scales is built using the Harris function,
$$\mu(\mathbf{x}, \sigma_I, \sigma_D) = \sigma_D^2 g(\sigma_I) * \nabla I(\mathbf{x}) \nabla^T I(\mathbf{x}) \tag{11}$$
where
$$\nabla I(\mathbf{x}) = (I_x(\mathbf{x}), I_y(\mathbf{x}))^T \tag{12}$$
is the spatial gradient of the image, σ_I an integration scale, and σ_D a differentiation scale. Initial salient points are then selected at the local maxima of each level of the representation.

2. an iterative algorithm is applied to simultaneously detect the location and scale of salient points. The extrema over scales of the Laplacian-of-Gaussian,
$$|\mathrm{LoG}(\mathbf{x}, \sigma_n)| = \sigma_n^2 |I_{xx}(\mathbf{x}, \sigma_n) + I_{yy}(\mathbf{x}, \sigma_n)| \tag{13}$$
is used to select the scale of salient points.

Since the algorithm does not produce a measure of saliency strength, equal weights were assigned to all resulting salient points.

The DiscSal saliency detector selects the salient features that are discriminant for the class of object against objects in the other classes, and then applies a biologically inspired model to detect salient points [13]. It was implemented as follows.

1. images are projected onto a K-dimensional feature space, and the marginal distribution of each feature response under each class $P_{X_k|Y}(x|i), i \in \{0, 1\}, k \in \{0, \ldots, K-1\}$, is estimated. The features are then sorted by descending marginal diversity,
$$\mathbf{md}(X_k) =< KL[P_{X_k|Y}(x|i)||P_{X_k}(x) >_Y \tag{14}$$
where $< f(i) >_Y = \sum_{i=1}^{M} P_Y(i) f(i)$, and $KL[p||q] = \int p(s) \log \frac{p(x)}{q(x)} dx$ the Kullback-Leibler divergence between p and q.

[3]The executable codes for implementing the two detectors are respectively available on-line at http://www.robots.ox.ac.uk/~vgg/research/affine/detectors.html and http://www.svcl.ucsd.edu/projects/saliency/.

2. features which are discriminant because they are informative about the background class ($Y = 0$) but not the class of interest ($Y = 1$), i.e. $H(X_k|Y = 1) < H(X_k|Y = 0)$ or that have too small energy to allow reliable inference, $Var(X_k) < T_v$, are eliminated.

3. the features of largest marginal diversity are selected.

4. a saliency map is generated by a biologically inspired architecture and salient points are determined by a non-maximum suppression stage which sets the scale of each salient location to the spatial support of the feature of largest response at that location.

The method is made scale adaptive by including in the candidate feature set the discrete cosine transform (DCT) features obtained by projecting, onto the 8x8 DCT basis, the result of a 4-level Gaussian pyramid image expansion.

The two sets of salient points were then fused into a Bayesian saliency (BayesSal) map according to (10), and finally, the centers of the resulting Gauss mixture were selected as salient points.

4.2 Salient locations

We start the evaluation of the Bayesian saliency detector by examining the salient locations detected for different object classes. Figure 7 presents some examples of the salient locations produced by the three detectors (locations with saliency strength lower than 50% of the largest are omitted). Note how Bayesian saliency combines the strengths of the two saliency mechanisms: while DiscSal forces HarrLap to focus in the area of the object of interest, the addition of HarrLap helps "clean up" some of the unstable locations detected by DiscSal. To obtain an objective characterization of the improvement, we measured the precision of the salient locations detected with the three methods. For this, the set of points on the saliency map of saliency strength greater than a threshold (equal to $Th_{sal}*$(maximum saliency strength), with $Th_{sal} \in \{0, 0.1, \ldots, 1\}$ on this experiment) was first selected, the number of points inside the ground truth (a manually produced bounding box of the object) was counted, and precision was measured as the ratio between the number of points inside the ground truth and the total number of selected points, i.e.

$$precision = \frac{\#\text{ of points inside the ground truth}}{\text{total }\#\text{ of points selected}}.$$

Finally, precision was over all images in the test set.

Figure 8 shows the measured precisions, as a function of the threshold Th_{sal}, on the face and motorbike classes (results on the other two classes were similar, and are omitted for brevity), with the three saliency detectors (and various values of σ for BayesSal). Several interesting observations can be made. First, HarrLap performed quite poorly, confirming the argument that DI detectors do not provide much information about the object of interest. Second, BayesSal with large σ tends to have the best precision, indicating that DD saliency is a crucial requirement for achieving informative salient points. Note, however, that even for BayesSal with small σ precision is

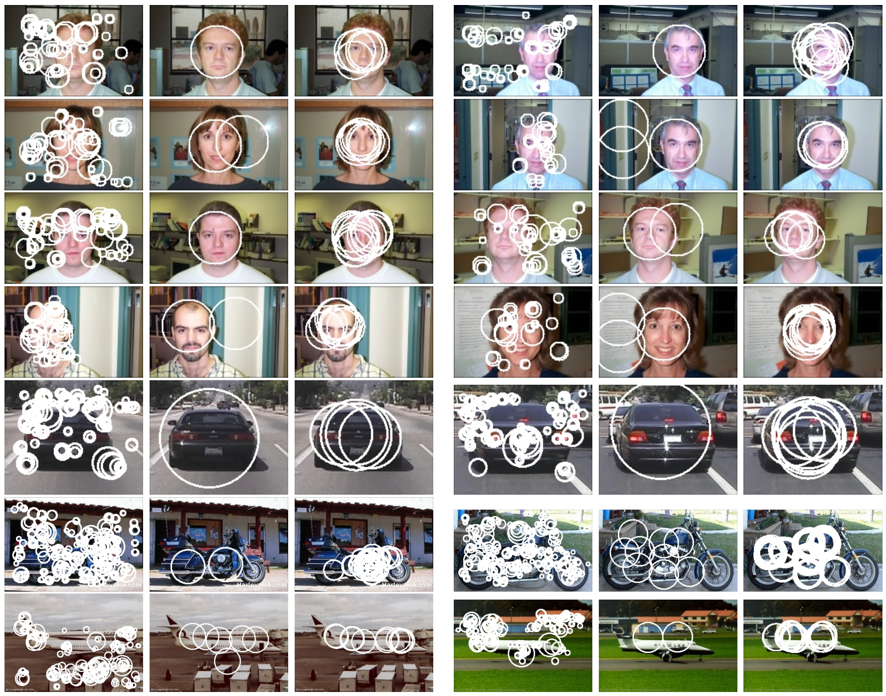


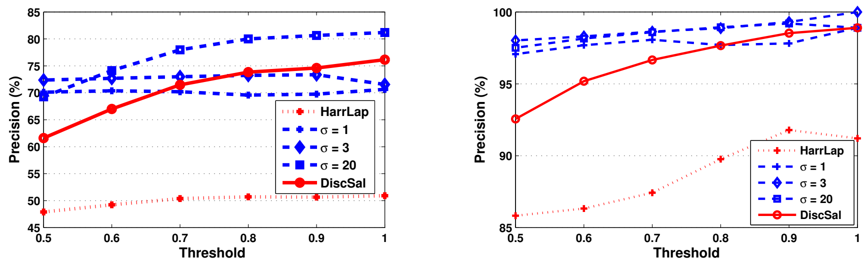
Figure 7: Examples of Bayesian saliency detection. (left) HarrLap, (middle) DiscSal and (right) BayesSal.

considerably higher than that of HarrLap. This is due to the *focus of attention* mechanism introduced by DD, which penalizes the DI points located far from the object of interest. Third, because when regularized by the more localized HarrLap saliency maps Bayesian saliency becomes much sparser than what is possible with DiscSal alone, the corresponding precision curves tend to be flatter, i.e. more insensitive to the threshold value. Finally, the best performance was always achieved with a saliency detector somewhere in between the DI (HarrLap) and DD (DiscSal) extremes, i.e. BayesSal with $\sigma \in (0, \infty)$. This confirms the argument that optimal performance requires a trade-off between the ability to meet application-specific goals (in this case "class-discrimination") and universal properties that assure good generalization (in this case "salient point stability", the criterion for which the Harris detector is optimal).

4.2.1 Segmentation of samples

To evaluate the robustness of BayesSal locations, the segmentation algorithm of Section 3 was applied to to two classes, face and rear-car, which can be well represented by a template. The following parameters were used: $k = 60$, $T_{corr} = 0.6$, $T_{in} = 12$, number of masks = 20, number of bad matches = 40, and $T_{mask} = 0.4$.

The quality of the segmented examples was evaluated by comparing them with ground truth data. The relative overlap between the segmented example (represented



(a) Face (b) Motorbike

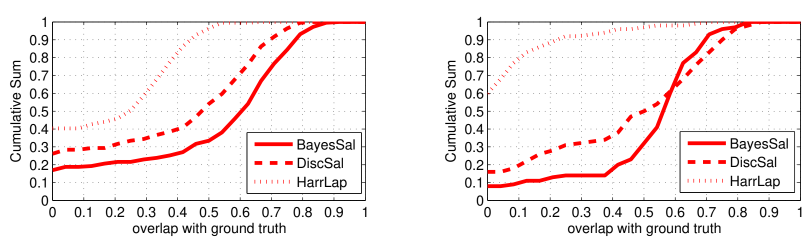
Figure 8: Precision of salient locations of the BayesSal (with various σ), DiscSal and HarrLap saliency detector.

as a rectangle on the image) and the ground truth was measured by

$$overlap(A, B) = \frac{|A \cap B|}{|A \cup B|} \tag{15}$$

where A, B are two bounding boxes and $|A|$ the area of A.

Figure 9 shows the cumulative distribution function for the relative overlap of the examples segmented by the three saliency detectors. Not surprisingly, BayesSal achieves the best performance, i.e. the curve closer to a delta function located at 100% overlap, confirming the advantages of relying on posterior distribution for salient locations that combines the discriminant power of DiscSal and the robustness of HarrLap. To provide a sense for the quality of the segmented patches, examples of faces segmented with various values of overlap are also shown in Figure 10. Interestingly, the algorithm produces "zero" overlap for about 10% of the images. Closer investigation of these cases reveals that they are images with either poor illumination or scale drastically different from the remainder of the examples (which falls outside the range of scales covered by our features), or could even be argued not to belong to the class (e.g. cartoons of faces). Examples of these "outliers" are shown in Figure 11.



(a) (b)

Figure 9: Cumulative distribution function for the relative overlap between segmented examples and manual ground truth for (a) faces and (b) cars.

0.5 0.6 0.7 0.8 0.9

Figure 10: Illustrative examples of segmented faces with overlap measures ranging from 0.5 to 0.9.

Figure 11: Images rejected by the segmentation algorithm.

4.3 Classification of saliency maps

Finally, we compared the performance of the different saliency detectors on an object detection task. In particular, we used the simple classifier proposed in [13], which consists of feeding a histogram of saliency map intensities to a support vector machine (SVM), and measuring the probability of classification error. This experiment quantifies how relevant the extracted saliency information is for recognition purposes. The performance of each classifier is measured by the receiver-operating characteristic (ROC) equal-error classification rate ($p(False\ positive) = 1 - p(True\ positive)$). As presented in Table 1, BayesSal generated better classification results than the two individual saliency detectors, DiscSal and HarrLap. For completeness, the table also presents the results, on this database, of a state-of-the-art method for recognition from cluttered scenes (the constellation-based classifier of [2]). Despite its simplicity, the saliency-based classifier achieves better recognition rates.

References

[1] S. Agarwal and D. Roth, "Learning a sparse representation for objection detection," In *Proc. ECCV 2002*, Vol. 4, pp. 113-130, 2002.

[2] R. Fergus, P. Perona and A. Zisserman. "Object Class Recognition by Unsupervised Scale-Invariant Learning," *Proc. IEEE Conf. on Computer Vision and Pattern Recognition,* 2003.

[3] E. Borenstein and S. Ullman "Learn to Segment," In *Proc. ECCV*, pp. 315-328, 2004

Dataset	BayesSal	DiscSal	HarrLap	constellation [2]
Faces	**98.5**	97.2	61.9	96.4
Motorbikes	**96.5**	96.3	74.8	92.5
Airplanes	**93.9**	93.0	80.2	90.2
Car Rear	**100.0**	100.0	92.7	90.3

Table 1: SVM classification accuracy based on different detectors.

[4] C. Harris and M. Stephens. "A combined corner and edge detector," *Alvey Vision Conference,* 1988.

[5] Förstner. "A framework for low level feature ex-traction," *Proc. of ECCV,* pp. 383-394, 1994.

[6] D. G. Lowe. "Object recognition from local scale-invariant features," *Proc. ICCV,* pp. 1150-1157, 1999.

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

[8] T. Kadir and M. Brady. "Scale, Saliency and Image Description," *Int'l. J. Comp. Vis.,* vol. 45, pp. 83-105, Nov. 2001.

[9] L. Itti, C. Koch and E. Niebur. "A model of saliency-based visual attention for rapid scene analysis," *IEEE Trans. PAMI,* 20(11), 1998.

[10] C. Privitera, L. Stark. "Algorithms for defining visual regions-of-interest: comparison with eye fixations," *IEEE Trans. Pattern Analysis & Machine Intelligence,* vol.22, Sept. 2000, pp.970-82.

[11] K. Walker, T.F. Cootes, and C.J. Taylor. "Locating Salient Object Features," *Proc. British Machine Vision Conf.,* pp. 557-566, 1998.

[12] B. Schiele and J. Crowley. "Where to look next and what to look for," *Intelligent Robots and Systems (IROS'96),* pp. 1249-1255, 1996.

[13] D. Gao and N. Vasconcelos. "Discriminant Saliency for Visual Recognition from Cluttered Scenes," *Proc. NIPS*, 2004

[14] K. Mikolajczyk and C. Schmid, "Scale and Affine invariant interest point detectors," *IJCV*, Vol. 1, No. 60, pp. 63-86, 2004.

[15] P. Viola and M. Jones. "Robust real-time object detection," 2^{nd} *Int. Workshop on Statistical and Computational Theories of Vision Modeling, Learning, Computing and Sampling,* July 2001.

[16] R.C. Gonzalez and R.E. Woods, *Digital Image Processing,* Addison-Wesley, Reading, MA, 1992

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

[18] M. A. Fischler, and R. C. Bolles, "Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography," *Communications of the ACM,* Vol. 24, No. 6, pp. 381-395, 1981.

[19] Gy. Dorko and C. Schmid "Selection of Scale-Invariant Parts for Object Class Recognition", *Proc. ICCV*, pp.634-640, 2003

[20] A. Opelt, M. Fussenegger, A. Pinz and P. Auer, "Weak Hypotheses and Boosting for Generic Object Detection and Recognition", *Proc. ECCV*, pp. 71-84, 2004

[21] G. Wahba, *Spline Models for Observational Data,* Soc for Industrial & Applied Math, 1990

[22] F. V. Jensen, *An Introduction to Bayesian Networks,* Springer, 1996

SVCL-TR 2005/01

June 2005

A Bayesian Architecture for Combining Saliency Detectors

Dashan Gao