Using Hidden Scale for Salient Object Detection

Bernard Chalmond, Benjamin Francesconi, and Stéphane Herbin

***Abstract*—This paper describes a method for detecting salient regions in remote-sensed images, based on scale and contrast interaction. We consider the focus on salient structures as the first stage of an object detection/recognition algorithm, where the salient regions are those likely to contain objects of interest. Salient objects are modeled as spatially localized and contrasted structures with any kind of shape or size. Their detection exploits a probabilistic mixture model that takes two series of multiscale features as input, one that is more sensitive to contrast information, and one that is able to select scale. The model combines them to classify each pixel in salient/nonsalient class, giving a binary segmentation of the image. The few parameters are learned with an EM-type algorithm.**

***Index Terms*—Focus, learning, object detection, probabilistic modeling, remote sensing, saliency, scale.**

I. Introduction

A. General Context

THE interpretation of remotely sensed images is faced with two kinds of complexity issues. The overwhelming amount of data generated by future systems raises a serious problem concerning their exploitation either by human or machines; the huge size of images imposes heavy computational constraints.

Indeed, in the near future, remote-sensing systems will be likely to transmit images of size up to several thousands of megapixels. Modern sensors will produce images with increasing resolution and field of view allowing new types of functions, in particular automatic target recognition (ATR).

The general context of this article is the detection of objects, typically all movable man-made objects, in satellite images of ground areas. This problem is difficult mainly because of the variety and inner complexity of the scenes to be processed (see Fig. 1 for an insight). In general, one cannot expect to have access to any usable model, neither for the background nor for the objects, and datasets are too poor compared with the variability of all possible situations, which makes learning methods based on a huge number of labeled samples not practicable. Algorithmic design is doomed to make the best use of *a priori* knowledge.

A global detection task can be organized in the following two steps.

1) **Focus**. This stage is devoted to the fast localization of *salient regions* using low computational cost algorithms. Salient regions indicate where an object is likely to stand, but have no precise semantical interpretation. The regions easily discriminated from an object, like flat areas, should be readily discarded. The process aims at a detection rate equal to 1, meaning that no object should be missed, while the false positive rate is allowed to remain relatively high. Note that this tradeoff is not usual in detection theory, where, generally, the false positive rate is fixed while the detection rate is the variable to be optimized.
2) **Detection/Recognition (DR)**. Once a large part of the image has been rejected as background, more time-consuming processing can be applied to smaller salient regions in order to make the final decision. The objective here is to reduce the number of false positives while keeping the detection rate as high as possible.

Fig. 1. Example of a remote sensed image (3200 × 3200 pixels, source: http://ortho.mit.edu/nsdi/draw-ortho.cgi?image=241898). The objects of main interest like planes are sparsely distributed among other objects (buildings, taxiways, embarkation ramps, etc.).

Focus distinguishes between two general classes, namely an *object of interest class*, containing for instance visual structures shared by every object like planes, cars, etc., and a *background class*. It is typically a bottom-up procedure, while DR may also include a top-down scheme exploiting more precise data models.

B. Overview of the Approach

In this paper, we propose a decision process designed to contribute to the focus stage in complex images, typically gray level aerial or satellite images. One of the difficulties to be solved is the handling of objects with multiple sizes and orientations surrounded by textured backgrounds. It is allowed to use few samples to estimate the algorithm parameters. Although the main

Manuscript received December 6, 2004; revised October 31, 2005. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Joachim M. Buhmann.

B. Chalmond is with the Centre de Mathématiques et de Leurs Applications, CNRS (UMR 8536), Ecole Normale Supérieure de Cachan, 94235 Cachan Cedex, France (e-mail: bernard.chalmond@cmla.ens-cachan.fr).

B. Francesconi and S. Herbin are with the Department of Information Processing and Modeling, ONERA, 92322 Châtillon Cedex, France (e-mail: benjamin.francesconi@onera.fr; stephane.herbin@onera.fr).

Digital Object Identifier 10.1109/TIP.2006.877380

aim of our approach is not computation time management, processing steps should been designed to be easily implemented in a parallel way.

The result of the algorithm is a binary segmentation of the image where only the "object" regions are worth being further processed in a DR phase. Local features are computed at each image site and used to attribute a class, "object" or "background."

Note that focus is not detection in the sense that no decision is taken about the true nature of the object; it is based on clues indicating that an object could be present or not.

Salient structures, i.e., objects, will be loosely defined as being *contrasted relatively to the background* and having *a bounded spatial extension and one or several closed contours*. In our target application, objects may also be noncompact, like planes, which brings additional difficulties. Hence, objects need to be described by at least two features: *scale and contrast*. These two features are intimately related: On the one hand, contrast is distributed in a region the extension of which depends on the local scale; on the other hand, contrast defines an area the size of which is related to what can be called scale. We think that these two notions cannot (and should not) be considered separately. In an image, scale is a *hidden parameter* that constrains local contrast spatial organization. Analyzing contrast is, therefore, the natural way to reach back a hidden scale.

In preliminary studies, we found it quite easy to design detectors for contrasted structures when their size was known. We were able to design filters, including a scale parameter that can detect quite well most contrasted structures, provided the scale parameter was correctly set (see Section II for an illustration). However, if one wants the detection to be fully automatic, the "appropriate" scale parameter has to be locally estimated. One basic idea is to compute another local feature from which one can extract some useful information about the spatial extension of the local structure, and use it to parametrize the local detectors.

The detector presented in this paper is a *multiscale saliency detector*. It uses a probabilistic model which achieves the interaction between multiscale contrast detectors and scale informative features. It can be considered as performing a weighted sum over the scales of the detector responses to take the final decision. The weights depend on the relative values of the scale features: the most likely scale is the most emphasized. The global detector is modeled as a mixture of logistic models, and uses few parameters that can be learned from image samples with an expectation-maximization (EM) type algorithm.

As a summary, the key point of our approach is to divide the focus task into two parts. One part consists of estimating, locally, a natural analysis scale based on local contrast elements. The other part uses a fixed scale contrast detector at this estimated scale to make a local decision. The main issues of the approach are the definition of contrast and scale sensitive local detectors and the design of a local detector combination scheme.

C. Related Work

Object detection and recognition are longstanding problems in computer vision, and are addressed in a wide range of applications: face detection, character recognition, vehicle detection, etc., [1]–[8]. However, the literature about object detection in aerial or remote-sensed images appears to be more limited. Indeed, the need of generic object models, the lack of a usable database, combined with demanding computational time requirements are strong limiting factors. Road extraction or building detection are connected remote sensing issues [9], [10]. The problem we address in this paper is the detection of objects on high resolution images (less than 1 m) without any specific knowledge of their structure.

Few studies have been concerned with computational efficiency issues. The power of coarse to fine strategies have been demonstrated in [1] and [2], while [7] uses boosting to select a cascade of classifiers. All those studies are based on statistical estimation and require the availability of a large learning database.

Object modeling has been widely worked out in the literature. However, most of the studies have addressed the problem of single object model and its use in matching. The most flexible and versatile approaches consider objects as a collection of parts [3]–[5], as an arrangement of simple features [2] or as a distribution of interest points on which locally invariant features are computed [11]–[13]. More recently, researchers have been interested in defining models of object categories for image retrieval applications [14]–[16]; their approach relies on a few highly resolved images sampling each category, which is a prerequisite not easily satisfiable in our target context. Some applications on traffic surveillance have developed very specific models for car detection in constrained setting [8].

Focus and saliency have been treated within different frameworks. While many studies [17]–[19] try to mimic human visual perception scheme (visual search, attentional phenomena), our work is guided by purely algorithmic and practical constraints. It is closer to the approaches of Kadir [20] and Lindeberg [21], in the sense that we want to detect some very general interesting elements in an image, and because scale is a central parameter. Kadir [20] also uses two multiresolution components. One is used for scale selection and measures feature-space unpredictability. The second one measures interscale unpredictability. The product of these two terms encodes what is called *scale saliency* at the selected scale. In our approach we aim at formalizing in a more coherent way the interaction between these two components.

Paper organization. Section II describes the elementary building blocks needed by the probabilistic mixture model described in Section III. In Section IV, we develop an EM-type algorithm used to learn the model parameters. Finally we illustrate our approach with experimental results in Section V. Section VI discusses several issues about scale correspondence between features.

II. Salient Structures and Scale

A. Salient Structure Features

Contrast features such as Gabor, derivative of Gaussian or Canny filters are standard tools of image processing. They are able to reveal local oriented patterns but are not designed to be sensitive to global extended spatial structures. Furthermore,

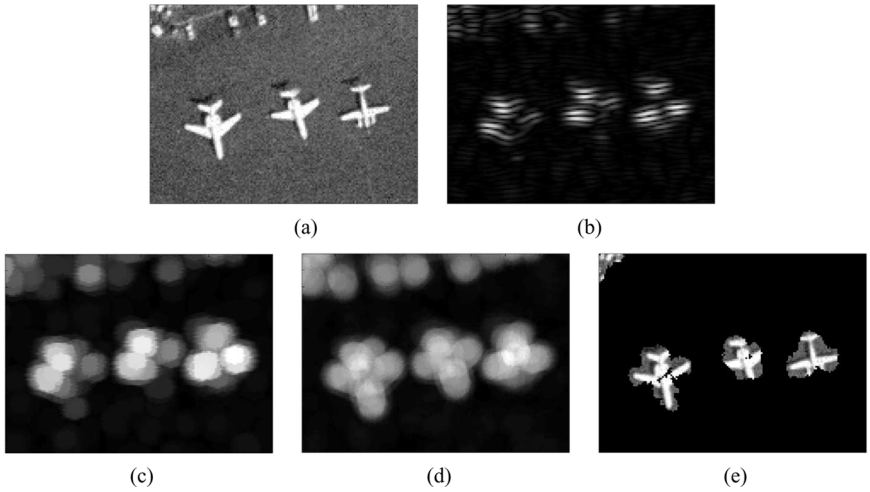


Fig. 2. Illustration of a salient structure detection for a given neighborhood width. The width of the planes is about 30 pixels. (a) Planes in an image. (b) Local edges generated by Gabor filter oriented at 0°. (c) Oriented dilated edges. (d) Sum of oriented dilated edges. (e) Result of the detection.

since the sought after objects of interest are essentially characterized, besides their contrast, by their spatial extension, there is a need to connect the decision process with a notion of scale.

There are several ways to introduce the information of scale in a detector. The technique we found useful consists of computing, at each site s, characteristics of local detectors over a spatial neighborhood V_s^e. The neighborhood width e acts as a *scale of analysis* aimed at revealing salient structures of corresponding size.

The proposed salient structure feature, referred as SaS in the rest of this paper, is based on local contrast features. We used Gabor filters in our experiments since they were proved to possess good signal analysis properties [17], [22]–[25] and rather low computational cost. Let $G_s(\alpha)$ be the absolute value of the imaginary part of oriented Gabor filters responses computed at pixel s with orientation α, typically 0°, 45°, 90°, and 135°. The standard deviation σ_{edge} of the Gabor filter gaussian component is fixed to a small value depending on the sharpness of image contours, which can be known *a priori* from sensor characteristics. The oscillating component parameter is set so that there is a single oscillation of the periodic component in a window of size $2\sigma_{\text{edge}}$. Such filters behave as local oriented edge detectors.

Let us consider a neighborhood V_s^e of width e centered at every pixel s. The parameter e is seen as a *scale parameter*. We define two SaS features, $D_s^e(1)$ and $D_s^e(2)$

$$D_s^e(1) = \sum_{\alpha} \max_{s' \in V_s^e} G_{s'}(\alpha) \tag{1}$$

$$D_s^e(2) = \sum_{\alpha} \sum_{s' \in V_s^e} G_{s'}(\alpha). \tag{2}$$

Feature (1) accumulates morphological dilations of oriented local edges computed for several orientations. The highest responses are obtained in neighborhoods containing all contrast orientations, i.e., neighborhoods potentially containing a closed contour bounding the salient structure. Feature (2) replaces the dilation step by a summation over all orientations in a neighborhood. The use of an averaging rather than a maximizing phase is expected to capture fuzzier contours and be sensitive to a high density of contours. It will be shown how to make the two detectors cooperate in a probabilistic mixture model in Section III.

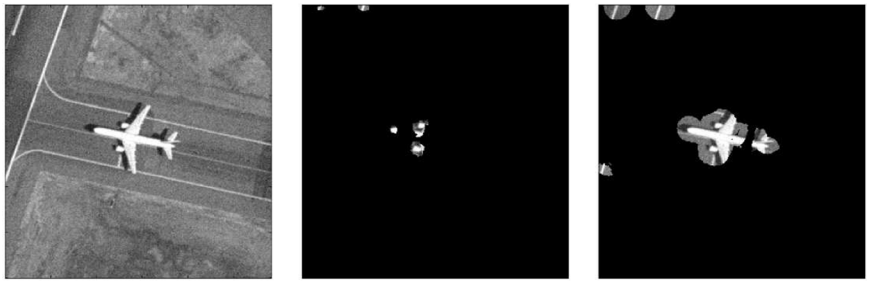
Let us illustrate the ability of SaS feature (1) to detect salient structures. Fig. 2 shows the sequence of operations achieving this detection. The first step is the computation of Gabor filters for various orientations [see Fig. 2(b) for orientation at 0° angle] followed by a neighborhood dilation [Fig. 2(c)]. The final SaS feature map is the sum of all the dilations for every orientation [Fig. 2(d)]. Salient regions are detected by thresholding the feature map [Fig. 2(e)]. In order to distinguish the feature from its exploitation, we will only talk of detection when a binary thresholding operation is involved.

Delocalizing the contrast detections is expected to provide the decision process with two interesting properties: an increase in informative power due to the local geometrical interactions it creates and a robustness to local changes due to the use of a wide spatial extension. The salient structures to be detected in the image are assumed to contain one or several components with closed and bounded contours, implying that contrast directions associated with the object of interest sample all orientations but with different rates. Therefore, one way to enhance saliency is to accumulate evidences of various oriented local contrasts located in a given neighborhood. This global accumulation ensures also an invariance to rotation since no specific orientation will be favored.

B. On Scale Parameter

The SaS features described above depend on the choice of the scale parameter e. This section illustrates the importance of this parameter.

Detection with fixed scale parameter. The same SaS detector, i.e., with the same scale parameter value e and threshold, is applied on images of similarly contrasted structures but of different size [Figs. 2(a) and 3(a)]. This value was first manually selected to give good results on images with small planes [Fig. 2(e)] and then used on images with big planes [Fig. 3(a)].



(a) (b) (c)

Fig. 3. Illustration of the importance of the scale parameter. This plane has a width of about 80 pixels. For this big plane, the appropriate scale parameter e is larger than the one for the small planes of Fig. 2. (a) Object of interest. (b) Result of detection using the same scale parameter as for small planes (Fig. 2). (c) Result of detection using an adapted scale parameter.

As we can see in Fig. 3(b), the results are degraded if the neighborhood is not consistent with the structure size. The appropriate scale parameter gives the results of Fig. 3(c). In general, in our experiments, we noted that even if the objects could be almost detected, a badly chosen scale parameter leads to an increase of false positives and/or to badly segmented object regions.

Detection with adapted scale parameter. The example above shows that scale information is crucial to obtain good detection. Ideally, the appropriate scale parameter should be estimated in each site so as to correctly parametrize the SaS detector. This can be achieved using scale sensitive features, referred to as SC in the following. As in Lindeberg [21], we exploit the fact that local extrema over scale e of a carefully chosen feature H_s^e computed locally at site s are likely candidates to correspond to interesting structures. It is expected that such an appropriate scale can help to choose an appropriate scale for SaS features.

Scale-space theory is a well-known formalism based on the suitable normalization of differential operators. Local entropy characteristics computed on varying size windows [20] also show interesting scale sensitivity phenomena. Following this way, the SC features H_s^e used in the experiments of Section V are based on the local entropy

$$E_s^e = -\sum_{z\in Z} p_s^e(z)\log p_s^e(z) \tag{3}$$

where $\{p_s^e(z), z\in Z\}$ is the empirical probability distribution of the gray levels $\{z, z\in Z\}$ computed over a circular window of radius e and centered in s. The discrete function $\{H_s^e,\ e\in\mathcal{E}\}$ is a weighted smooth version of $\{E_s^e,\ e\in\mathcal{E}\}$ as it follows. Let w_s^e be the $L1$ norm between the two distributions $\{p_s^e(z), z\in Z\}$ and $\{p_s^{e+1}(z), z\in Z\}$. Then, the SaS feature is defined as

$$H_s^e = \widetilde{E}_s^e \widetilde{w}_s^e \tag{4}$$

where $\left\{\widetilde{E}_s^e,\ e\in\mathcal{E}\right\}$ denotes a smoothed version of $\{E_s^e,\ e\in\mathcal{E}\}$. $\mathcal{E}$ is the set of possible scales. Smoothing is used to clean the signal around the maximum value of $\{H_s^e,\ e\in\mathcal{E}\}$ on salient sites. When such a peak is present, its abscissa e can be used to estimate the local scale at site s.

Section III presents a framework for combining SaS and SC features in a unifying probabilistic framework. Both are controlled by a common parameter scale.

III. Probabilistic Modeling

We propose a probabilistic model which combines multiscale features for scale selection and saliency detection. Our goal is to make pixel-wise detection on a sampled image: Each pixel s will be assigned a *saliency flag* (1 or 0) indicating whether it belongs to an object or not, thus, obtaining a binary segmentation of the image. In this probabilistic model, both scale and saliency flag are considered as random variables. The output, at pixel s, is the probability that this pixel belongs to an object or not.

A. Mixture Model

In order to present a more formal description of the probabilistic model, let us introduce several notations.

- $s\in\mathcal{S}$: Pixel location in the image sampled on a grid $\mathcal{S}$.
- ϵ_s: Scale random variable at position s, taking values e in $\mathcal{E}$, the set of possible scales.
- A_s: Binary random variable indicating object detection at pixel s. $A_s=1$ if s belongs to a salient region, $A_s=0$, otherwise. It represents the final decision we aim at.
- A_s^e: Binary random variable indicating detection at pixel s and fixed scale e.
- H_s^e (scalar), for $e\in\mathcal{E}$, local SC features.
- $\mathbf{D}_s^e = (D_s^e(1)\dots D_s^e(C))^T$, for $e\in\mathcal{E}$, local SaS features. The scalar C is the number of components used. In our experiments, $C=2$.
- $H_s = \{H_s^e, e\in\mathcal{E}\}$ the collection of scale features.
- $Y_s = \{(\mathbf{D}_s^e, H_s^e), e\in\mathcal{E}\}$ the collection of input data.

We want to express the probability that pixel s belongs to an object $(A_s=1)$ or not $(A_s=0)$ given the whole multiscale local features Y_s. Introducing the "hidden" local scale random variable ϵ_s and using Bayes law, one has

$$\begin{aligned} P(A_s=a|Y_s) &= \sum_{e\in\mathcal{E}} P(A_s=a,\epsilon_s=e|Y_s) \\ &= \sum_{e\in\mathcal{E}} P(A_s=a|Y_s,\epsilon_s=e)P(\epsilon_s=e|Y_s). \end{aligned} \tag{5}$$

We make the hypothesis that ϵ_s only depends on H_s, and A_s only depends on $\mathbf{D}_s^e$ when ϵ_s is given. This fundamental hypothesis allows to separate contrast and scale components. Under these hypotheses $P(A_s = a|Y_s, \epsilon_s = e)$ reduces to $P\left(A_s = a|\mathbf{D}_s^e\right)$ and $P(\epsilon_s = e|Y_s)$ to $P(\epsilon_s = e|H_s)$. We also assume that when scale e is known, the global detector A_s is the same as the one given by the corresponding fixed-scale detector A_s^e. Equation (5) becomes

$$P(A_s = a|Y_s) = \sum_{e\in\mathcal{E}} P\left(A_s^e = a|\mathbf{D}_s^e\right) P(\epsilon_s = e|H_s).$$

By its structure, this model puts two components into interaction: local scale indicators and local saliency detectors. The conditional probability $P(\epsilon_s = e|H_s)$ acts as a weighting function on the responses of the fixed scale detectors A_s^e involved in the final detection. Ideally, when $P(\epsilon_s = e|H_s)$ is peaked (i.e., scale is perfectly known), the model comes down to select the appropriate saliency detector among detectors $\{A_s^e|\mathbf{D}_s^e,\ e \in \mathcal{E}\}$. In this case, final decision depends only on $\mathbf{D}_s^e$ at the selected scale. In the general case, interaction is a little more intricate since every fixed scale detector A_s^e contributes to the final decision depending on its SC feature H_s^e.

The final simplest decision is then made by comparing the output probability to a detection threshold T

$$\text{if } P(A_s = 1|Y_s) > T, \text{ then s belongs to an object,}$$
$$\text{otherwise it belongs to the background.}$$

B. Probability Models for Saliency Flag and Hidden Scale

The conditional probabilities are modeled as logistic functions [26]. Further notations are needed to complete the model. Let

$$\tilde{\mathbf{D}}_s^e = (1, D_s^e(1), D_s^e(2) \ldots D_s^e(C))^T = \left(1, \mathbf{D}_s^{eT}\right)^T$$

be an augmented vector of multiscale SaS features and

$$\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_C)^T$$

be the logistic coefficients. With these notations the logistic models are defined as

$$\begin{aligned} P_{\boldsymbol{\beta}}^{s,e}(a) &\doteq P\left(A_s^e = a|\mathbf{D}_s^e\right) \\ &= \frac{\exp\left[\left(\beta_0 + \sum_c \beta_c D_s^e(c)\right) a\right]}{1 + \exp\left(\beta_0 + \sum_c \beta_c D_s^e(c)\right)} \\ &= \frac{\exp\left[\left(\boldsymbol{\beta}^T \tilde{\mathbf{D}}_s^e\right) a\right]}{1 + \exp\left(\boldsymbol{\beta}^T \tilde{\mathbf{D}}_s^e\right)} \end{aligned} \tag{6}$$

and

$$P_\lambda^s(e) \doteq P(\epsilon_s = e|H_s) = \frac{\exp\left(\lambda H_s^e\right)}{\sum_{e'} \exp\left(\lambda H_s^{e'}\right)} \tag{7}$$

where λ is a positive scalar parameter. The choice of the logistic model as in [22] is motivated by its well-known ability to perform classification. It also allows learning using an EM formulation. The possibility of adding as many contrast features $D_s^e(c)$ as we want, so as to refine salient regions description, and the possibility of computing the multiscale features in a parallel architecture are other important advantages.

The mixture model fulfills the role we have anticipated: combine SaS and SC features in a unifying framework.

Role of $P_{\boldsymbol{\beta}}^{s,e}(a)$. The logistic model $P_{\boldsymbol{\beta}}^{s,e}(a)$ makes a linear combination of the input vector components $D_s^e(c), c = 1 \ldots C$ and compares the result to a threshold $(-\beta_0)$ to give a probability. Comparing this probability to 0.5 to decide which class the sample belongs to is equivalent to comparing $\boldsymbol{\beta}^T \tilde{\mathbf{D}}_s^e$ to 0. Somehow, the model splits the input space in two classes, with a hyperplane of equation $\beta_0 + \sum_c \beta_c D_s^e(c) = 0$. The coefficients $\beta_1, \ldots, \beta_C$ adjust the model sensitivity to each component $D_s^e(c)$ around the hyperplane. The SaS features $D_s^e(c)$, $c = 1 \ldots C$ have to be chosen so that they take high values for interesting saliency: the higher the feature value, compared with the threshold, the higher the probability of belonging to an object.

Role of $P_\lambda^s(e)$. The component $P_\lambda^s(e)$ of the mixture model is intended to make scale selection at pixel s, assigning to each scale $e \in \mathcal{E}$ a probability based on the SC features H_s^e. The probability $P_\lambda^s(e)$ can be rewritten as

$$P_\lambda^s(e) = \frac{1}{1 + \sum_{e' \neq e} \exp(\lambda(H_s^{e'} - H_s^e))}$$

showing that the probability of scale e only depends on the difference between H_s^e and $H_s^{e'}$, $e' \neq e$. Assuming $\lambda > 0$, the probability is highest at the scale for which H_s^e is maximum. The logistic model emphasizes the scales where H_s^e is high compared with $H_s^{e'}$, $e' \neq e$. The parameter λ adjusts the sensitivity of this enhancement. If the feature H_s^e reaches a marked peak over scale e, *which defines appropriate local scale*, the corresponding scale will have a high probability. This part of the model is devoted to the task of scale selection.

IV. Learning

The mixture model depends on the parameters $\boldsymbol{\theta} = (\boldsymbol{\beta}, \lambda)$. Its estimation can be done following a maximum likelihood principle. A straightforward optimization is not easy, but the introduction of "hidden" variable ϵ_s makes it feasible using an EM algorithm ([27] among many others). The learning is supervised, and we assume that we hold a series of N labeled feature samples $(Y_n, a_n), n = 1 \ldots N$ extracted from some image data representative of the problem.

Let us introduce several new notations.

- $\boldsymbol{\theta} = (\boldsymbol{\beta}, \lambda)$ the model parameters we want to estimate.

- $\mathbf{a} = (a_n)$, $n \in \mathcal{L} = 1 \ldots N$ the series of labels (1 for "salient," 0 for "non salient"), coming from a manual segmentation of a learning image, for instance.
- $P^n_{\boldsymbol{\theta}}(a) \doteq P(A_n = a|Y_n)$ the likelihood of each sample.
- $\boldsymbol{\epsilon} = (\epsilon_n), n \in \mathcal{L}$ and $\mathbf{e} = (e_n)$, $n \in \mathcal{L}$, the random vector of hidden scale for each sample and its realization, respectively.
- $P^n_{\boldsymbol{\theta}}(a, e) \doteq P(A_n = a, \epsilon_n = e|Y_n)$ the conditional joint probability of A_n and ϵ_n.
- $P_{\boldsymbol{\theta}}(\mathbf{a}, \mathbf{e}) = \prod_{n\in\mathcal{L}} P^n_{\boldsymbol{\theta}}(a_n, e_n)$ the likelihood for the joint probability law under independence hypothesis of the samples.

The likelihood we want to maximize over $\boldsymbol{\theta}$ is

$$P_{\boldsymbol{\theta}}(\mathbf{a}) = \prod_n P^n_{\boldsymbol{\theta}}(a_n). \tag{8}$$

With the EM algorithm, a local maximum of this quantity can be reached by maximizing iteratively over $\boldsymbol{\theta}$

$$F_{\boldsymbol{\theta}^{(k)}}(\boldsymbol{\theta}) = \mathbb{E}_{\boldsymbol{\epsilon}|\mathbf{a},\boldsymbol{\theta}^{(k)}}[\log P_{\boldsymbol{\theta}}(\mathbf{a}, \mathbf{e})] \tag{9}$$

where $\mathbb{E}_{\boldsymbol{\epsilon}|\mathbf{a},\boldsymbol{\theta}^{(k)}}$ is the conditional expectation with respect to $\boldsymbol{\epsilon}$, knowing $\mathbf{a}$ and $\boldsymbol{\theta}^{(k)}$, and where $\boldsymbol{\theta}^{(k)}$ is the estimated parameter at the kth iteration. Iterations on $\boldsymbol{\theta}^{(k)}$ are driven by the recursion

$$\boldsymbol{\theta}^{(k+1)} = \arg\max_{\boldsymbol{\theta}} F_{\boldsymbol{\theta}^{(k)}}(\boldsymbol{\theta}). \tag{10}$$

Computing the quantity $F_{\boldsymbol{\theta}^{(k)}}(\boldsymbol{\theta})$ (9) constitutes the E-step of the EM algorithm, while maximizing (10) is the M-step. Developing (9), and assuming the random hidden scales ϵ_n, $n \in \mathcal{L}$ are independent variables, it becomes

$$\begin{aligned} F_{\boldsymbol{\theta}^{(k)}}(\boldsymbol{\theta}) &= \sum_{n\in\mathcal{L}} \mathbb{E}_{\epsilon_n|a_n,\boldsymbol{\theta}^{(k)}}\left[\log P^n_{\boldsymbol{\theta}}(a_n, e_n)\right] \\ &= \sum_{n\in\mathcal{L}} \sum_{e\in\mathcal{E}} \frac{P^n_{\boldsymbol{\theta}^{(k)}}(a_n, e)}{\sum_{e'\in\mathcal{E}} P^n_{\boldsymbol{\theta}^{(k)}}(a_n, e')} \log P^n_{\boldsymbol{\theta}}(a_n, e). \end{aligned} \tag{11}$$

In this expression, everything is known, except the parameter $\boldsymbol{\theta}$. We use the Newton–Raphson algorithm in a second loop to realize the M-step, that is to maximize this expression over $\boldsymbol{\theta}$. It is also an iterative algorithm. First, initialize the parameter with $\boldsymbol{\theta}_{(0)}$. Then use iteratively the updating formula (12) until convergence. The ith iteration of the Newton–Raphson algorithm is described as

$$\boldsymbol{\theta}_{(i)} = \boldsymbol{\theta}_{(i-1)} - H^{-1}_{i-1} \nabla_{i-1} \tag{12}$$

where

$$\begin{aligned} H_{i-1} &= D^2_{\boldsymbol{\theta}} F_{\boldsymbol{\theta}^{(k)}}(\boldsymbol{\theta}_{(i-1)}) \\ \nabla_{i-1} &= \nabla_{\boldsymbol{\theta}} F_{\boldsymbol{\theta}^{(k)}}(\boldsymbol{\theta}_{(i-1)}). \end{aligned}$$

Problem: Maximize the likelihood:

$$P_{\boldsymbol{\theta}}(\mathbf{a}|\{Y_n, n \in \mathcal{L}\}) \doteq P_{\boldsymbol{\theta}}(\mathbf{a}) = \prod_n P^n_{\boldsymbol{\theta}}(a_n)$$

First loop: Iterative EM algorithm

- Initialization: $\boldsymbol{\theta}^{(0)}$
- **E step** (Expectation). Compute $F_{\boldsymbol{\theta}^{(k)}}(\boldsymbol{\theta})$ with (Eq. 11).

  Second loop: Maximization using Newton-Raphson algorithm
  - Initialization: $\boldsymbol{\theta}_{(0)}$
  - Compute iteratively $\boldsymbol{\theta}_{(i)}$ from $\boldsymbol{\theta}_{(i-1)}$ with (Eq. 12).

Fig. 4. Scheme of the learning procedure.

The quantity H_{i-1} is the Hessian of $F_{\boldsymbol{\theta}^{(k)}}(\boldsymbol{\theta})$ computed at $\boldsymbol{\theta}_{(i-1)}$ and ∇_{i-1} is the gradient computed at $\boldsymbol{\theta}_{(i-1)}$. When convergence is achieved on $\boldsymbol{\theta}_{(i)}$, the E-step is initiated with this new value for $\boldsymbol{\theta}^{(k+1)}$. As a summary, the learning scheme consists of two embedded loops, the EM loop containing the Newton–Raphson loop in the M-step. See Fig. 4 for a summary of all the stages of the EM learning scheme. Note that EM and Newton–Raphson algorithms only ensure that a local optimum is found.

V. Experimental Results

The approach described in the present study is a focus strategy aiming at pointing potential objects of interest, with a high detection ratio. It is designed to be followed by a further recognition step able to exploit more precise object models and should not be considered final. The features exploited—Gabor filter responses—are simple and not specific to a precise class of object. The learning phase is devoted to adapt the few model parameters to the type of image processed.

This section gives some insights on the algorithm performances on two different contexts: plane detection on an airport aerial image,[1] and detection of cars observed from their side in an urban environment.[2]

Several studies have been evaluated on the UIUC car database [4], [28], [29]. They describe a complete detection procedures and exploit a very specific image encoding and an object model learned on an extensive set of samples. They address a different problem than ours since the output of this type of algorithm is a series of true object locations and sizes. In a focus algorithm, we are essentially interested in finding the locations of salient objects. Their precise size or bounding region is a secondary information, and was not intended to be estimated as an output of the probabilistic mixture model (5).

[1]http://ortho.mit.edu/nsdi/draw-ortho.cgi?image=241898 and http://ortho.mit.edu/nsdi/draw-ortho.cgi?image=241902

[2]http://www.pascal-network.org/challenges/VOC/download/uiuc.tgz

Fig. 5. Typical extract of a scene containing planes close to a hangar (size is 773×626). Other contexts are isolated objects [Figs. 2(a) or 3(a)] or objects close to terminals (Fig. 8).

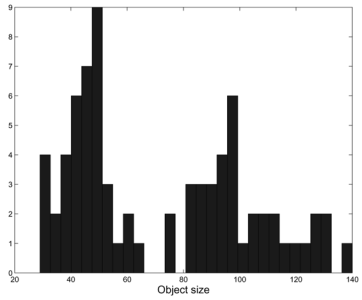


Fig. 6. Repartition of object sizes in the tested image. The ratio between the largest and the smallest object is about 5.

A. Performance Measures

Given the difference between the objectives of a focusing algorithm and a complete detection procedure, we propose to evaluate our approach on two types of performance measures. Both are based on ground truth regions centered around objects of interest.

The first one gives a hint of the amount of discrimination capacity still needed to complete a global scene interpretation starting from a focus output. A detection, i.e., a location proposed by the focusing algorithm, is assumed to be good if it is contained in a ground truth area closely bounding each object. To avoid multiple counting, each ground truth region can only be associated with one detection, the others being counted as false positives. This first type of measure will be referenced as *Focus* in the following.

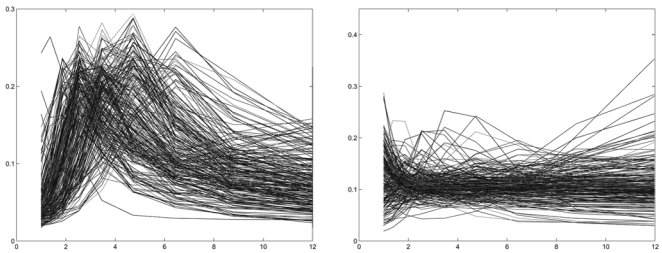
In the second type of measure, the algorithm outputs a series of regions bounding the object as tight as possible. We follow the criterion described in [28] to assess the validity of a given region based on region width and centroid position. This second type of measure will be referenced as *Region* in the following.

As it is now customary in object-detection problems, global performances are based on precision-recall points computed for various values of the control parameters or thresholds. They count the relative frequency of good and false decisions, regions, or locations. *Precision* is defined as the ratio of true positive decisions (TP) over all decisions $(TP + FP)$

$$\text{Precision} = \frac{TP}{TP + FP} \tag{13}$$

Fig. 7. Five patches used as positive examples. Objects have sizes from 30 to 40 pixels.

Fig. 8. Position of the five positive examples in the learning image (highlighted areas).



(a) (b)

Fig. 9. Distribution $\{P_\lambda^s(e),\ e \in \mathcal{E}\}$ ((7)) for each of the 241 learning samples. (a) Positive samples. (b) Negative samples.

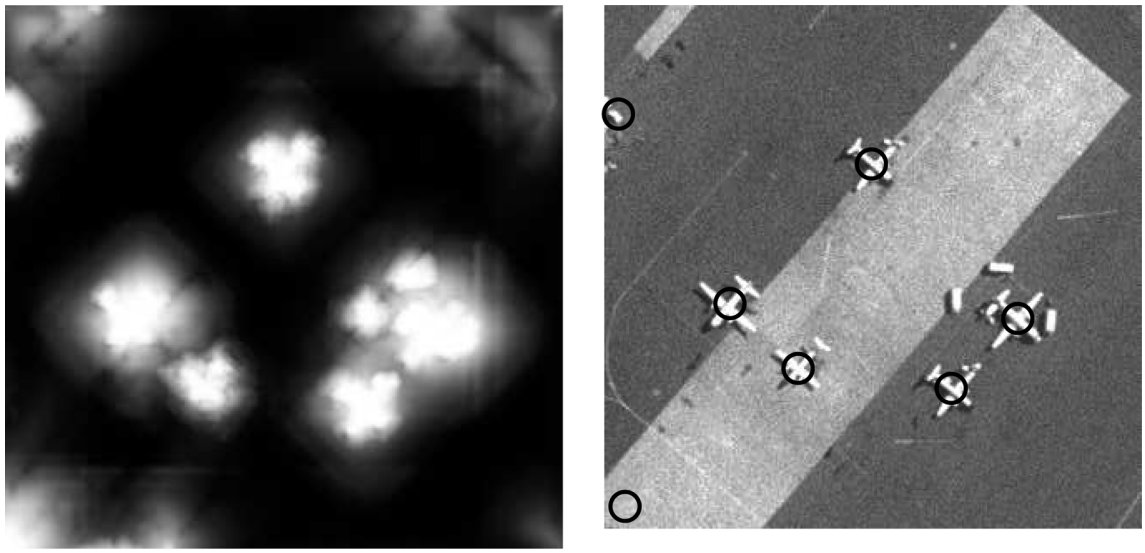
while *Recall* is the ratio of true positives over the number of expected good decisions (nP)

$$\text{Recall} = \frac{TP}{nP}. \tag{14}$$

We also give the values of missed objects $nP - TP$ and false alarms FP as performance indicators.

B. Plane Detection

In this first experiment, the focusing algorithm is tested on an aerial image of an airport (Fig. 1). From the global ortho-image, 14 subimages having each a resolution of 0.5 m were extracted and covering all the interesting areas in a total of $3778030 \approx 1943^2$ pixels. The images contain 75 planes having sizes with rather large variations (see Fig. 6) and with no favored orientation. An airport scene is complex and contains scattered and unevenly distributed sources of false alarms such as small



(a) (b)

Fig. 10. Illustration of salient object detection (threshold = .6, sigma = 15). Planes of interest have sizes from 35 to 45 pixels. Image size is 315 × 323. (a) Combined probability map. (b) Detected objects.

buildings, vehicles and embarkation ramps. The 75 planes on the airport scene can be qualitatively divided in 9 on taxiways [Figs. 2(a) or 3(a)], 26 close to hangars (Fig. 5) and 40 close to terminals (Fig. 8), this last category being considered as the most difficult *a priori*.

The mixture model assigns a probability of saliency $P(A_s = 1|Y_s)$ to each pixel s. Potential object locations are isolated from this map by applying a Gaussian smoothing filter and detecting local maxima. The value of each maximum is then compared to a detection threshold.

In order to deal with a wide object size interval, the probability map associated with each image has been computed by combining intermediate maps in a pyramid of subsampled images. A combination law associates the maximal probability in each subsampled probability map to each pixel in the image of highest resolution.

In the experiment, we used the features $\mathbf{D}_s^e = (D_s^e(1), D_s^e(2))^T$ and H_s^e presented in Section II (1), (2), (4). For SaS feature computation, the standard deviation of the Gabor filters is $\sigma_{\text{edge}} = 3$. The model has four parameters: $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2)$ and λ. The hidden scale e varies in a logarithmic scale from 1 to 12 pixels, and three sub-sampling steps are exploited to build the combined probability map.

Parameter estimation uses samples drawn from positive—the objects of interest—and negative regions in the image. Fig. 7 shows the five positive patches used in the experiment, all belonging to the same image (Fig. 8). In order to help learning, we selected from these positive regions only sample sites that showed a marked maximum in the scale selection signal. We selected the positive samples according to the value of their second order derivative, the 30% lowest. Negative samples were randomly chosen in the rest of the image so that positive and negative populations should be balanced. Note that after these selections, very few positive samples—241 sites in this experiment—are needed for learning, which makes parameter estimation very fast. The resulting estimated parameter values, used throughout the experiments, were $\boldsymbol{\beta}_0 = -10.1$, $\boldsymbol{\beta}_1 = 0.25$, $\boldsymbol{\beta}_2 = 1.18$, and $\lambda = 8.3$. Fig. 9 shows the distribution of the scale selection probabilities for the positive and negative samples after learning.

The role of the focusing algorithm is to provide a series of regions of interest to a further interpretation phase, either machine or human based. Ground truth regions are exclusively centered around planes [Fig. 11(a)], and do not include objects which may have similar features and sizes such as vehicles or embarkation ramps in an airport. The role of this very strict ground truth is to define a detection problem which may consist of an operational requirement.

Given the learned model, evaluation consisted of applying the global detection procedure, that is: probability map computation for a series of subsampled images, map combination using a maximum operator, smoothing, local maximum detection, and thresholding.

Fig. 10 shows the results of the focusing algorithm on a simple case where objects are well separated. Using a probability threshold of 0.6 and a Gaussian smoothing kernel of width 15, all the planes are detected.

Fig. 11 shows the detection results on a more complex scene for various thresholds and Gaussian kernel widths. Here, several low or unevenly contrasted planes are only locally detected by a part of their extension [Fig. 11(d)]. The use of a larger Gaussian kernel width lowers the number of local maxima per object at the expense of several missed detections [Fig. 11(e)].

Experimental results are shown in Fig. 12 as precision-recall points. Since we used two control parameters (thresholds and kernel width), results are represented as a collection of curves. In the target application, we are especially interested in detecting all the objects of interest while tolerating false positives that will be rejected in a further step. Table I shows the repartition of detections for various local maximum thresholds and a kernel

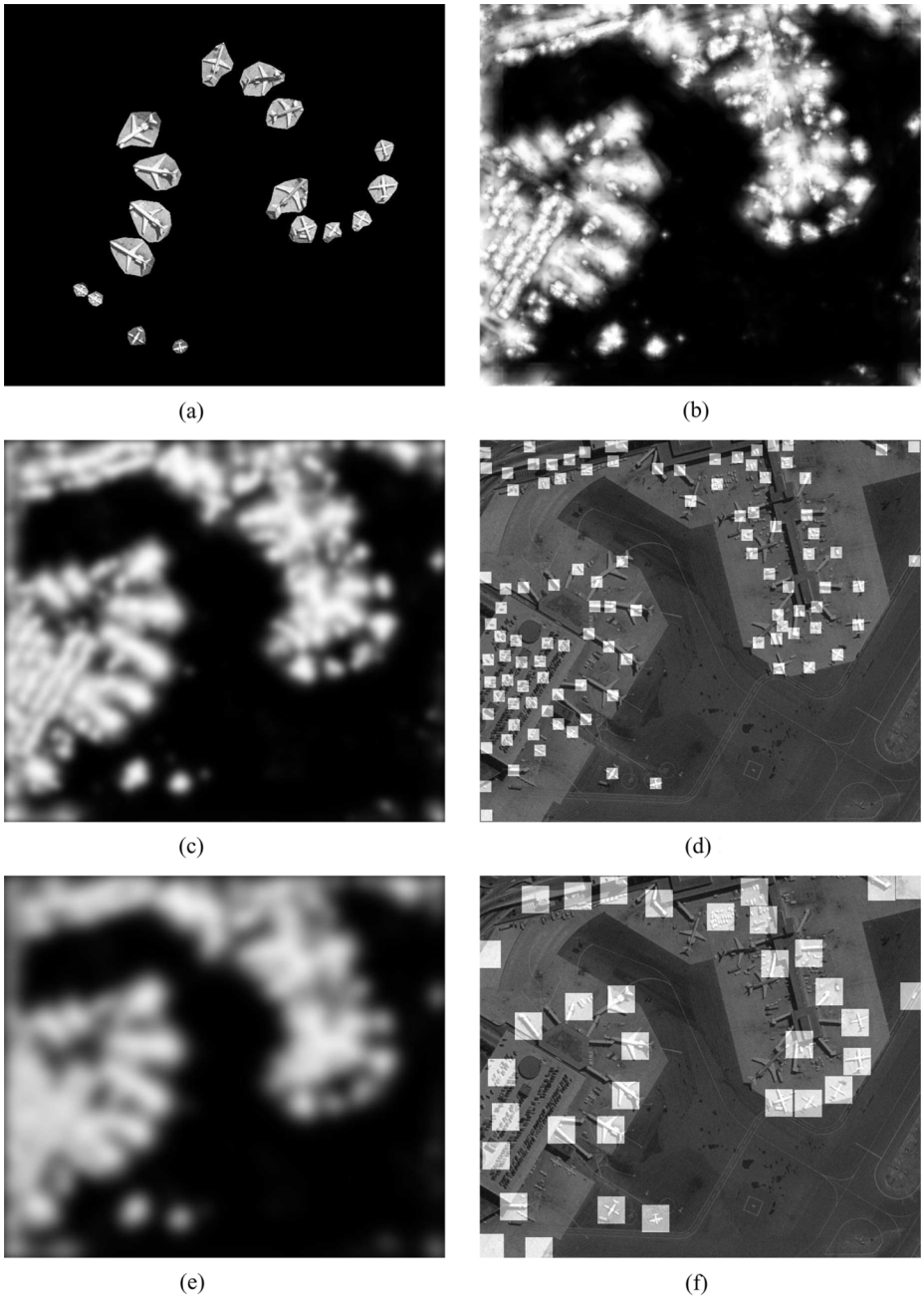


Fig. 11. Illustration of salient object detection for various thresholds and smoothing parameters. Planes of interest in this image have sizes from 30 to 90 pixels. Highlighted areas are associated with local maxima position. Image size is 884×766. (a) Ground truth regions. (b) Combined probability map. (c) Filtered probability map (sigma $= 15$). (d) Detected local maxima (threshold $= .6$). (e) Filtered probability map (sigma $= 30$). (f) Detected local maxima (threshold $= .3$).

of width 15 pixels. All the planes are detected for a precision of 15.8% corresponding to 400 false alarms to be rejected in a further phase. The total computation time is 620 seconds with an implementation in Matlab running on a Pentium IV at 2.4 Mhz, which corresponds to 0.2 ms per pixel. 70% of the time is spent on the computation of the SaS feature, the weighted entropy, which is not optimized in our implementation.

C. Car Detection

This context is quite different from the previous case since images consist of scenes where objects occupy a rather wide portion of the field of view and have smaller variations of size. The objects are side views of cars and present the same orientation and aspect. The background is in general highly textured, and objects have rather large illumination variations.

The data base contains learning and test sets. The test set is divided in two parts: 170 images containing 200 side views of cars with common size (100×40), and a second test set of 108 images containing 139 cars at various sizes with a ratio between the largest and smallest car of about 2.5.

The SC feature used in this experiment are Gabor filters with standard deviation $\sigma_{\text{edge}} = 3$ computed in four directions (0°, 45°, 90°, and 135 °). Instead of averaging the four directions as in the plane detection problem before performing dilation or spatial averaging (1), (2), we keep each direction independent.

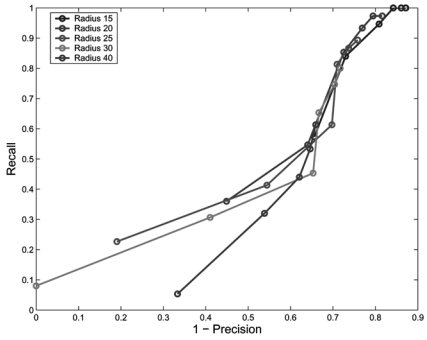


Fig. 12. Precision/recall points for various control parameters (threshold and spatial averaging). The linked points correspond to the same kernel width, but with different local maximum thresholds.

TABLE I
REPARTITION OF DETECTIONS FOR VARIOUS THRESHOLDS AND A SPATIAL AVERAGING OF WIDTH 15. VALIDATION CRITERION IS *FOCUS* (SEE TEXT FOR DEFINITION)

Threshold	**0.7**	**0.8**	**0.9**	**0.95**
FP	400	300	170	82
nP - TP	0	4	12	32
Recall	100.0%	94.7%	84.0%	57.3%
Precision	15.8%	19.1%	27.0%	34.4%

This structure is expected to catch the anisotropy of the signal since cars have clearly a favored orientation in the images. The SaS component (weighted entropy) remains unchanged, making the total number of parameters to be estimated equal to 10. The hidden scale e varies linearly from 1 to 20 pixels.

For learning, we used the first five and first ten images from the positive and negative samples of the training data base. This is a low number of learning samples compared with the $500{+}500$ images needed in [28]. The same mixture model parameters are used on both test sets.

The resulting saliency map $P(A_s = 1|Y_s)$ is exploited in two directions: as a focusing map or as an intermediate state in a complete object detection procedure. In the first case, focus points are detected after Gaussian smoothing with standard deviation of 15 pixels followed by local maximum search. In the second case, two different strategies are used, depending on the test set.

In test set I, all objects have roughly the same size (100×40 pixels). The saliency map is exploited by detecting the best salient windows of fixed size (80×40 pixels in our experimentation), which is equivalent to convolve the map with a uniform window of that size, and search for local maxima in the resulting filtered map (Fig. 13).

In test set II, objects have various sizes, and require a multiscale strategy to isolate their shape. We use a scale-space approach for blob detection ([21]) consisting of computing suitably normalized Laplacian of Gaussian filters of increasing standard deviation. Scale-space local maxima are sought as candidates for object location and size.

In both test sets, a detection candidate is counted as good if its estimated object location and size belong to a given uncertainty ellipse. We use the protocol and tolerance values as in [28]. Illustrations of the rigorousness of the acceptance criterion are presented Figs. 14 and 15.

Focus results are presented on Tables II and III for the two test sets. All objects are detected with a number of false alarms equal to 519 for test set I and 689 for test set II, corresponding to precisions of 27.8% and 16.8%.

Object detection results are presented on Tables IV and V for the two test sets.

Results for the single size case are comparable with those of [28] (see Table VI). Fig. 15 shows typical rejections of candidates for car detection. Missed detections happen mainly in low contrasted areas where the Gabor filters have weak responses: The SaS features are very sensitive to contrast densities, and were designed to behave so, since our target application is plane detection in aerial images.

Another source of misdetection is a distribution of local contrasts in background similar to those in the object shape. In these situations, the mixture model assigns saliency probabilities with no marked maximum, making object detection, i.e., filtering and local maximum search, unable to correctly select the accurate object location. This phenomenon is especially noticeable in test set II where the object detection procedure seldom finds the correct object size (Table V). The best achieved detection ratio in this test set is 37.4% for a precision of 11.3%.

Computation time was 0.3 ms per pixel on test set I, where 70% of the time is spent on SaS feature computation. The multiscale blob detection used in test set II adds 0.2 ms per pixel. On an image of size 200×150, total computation time is 9 s on test set I, and 15 s on test set II. Coding is in Matlab script running on a Pentium IV at 2.4 Ghz.

VI. DISCUSSION

A. Scale Correspondence

Scale is a concept that can have two interpretations: a) as a characteristic referring to the extension measure, i.e., size, of some physical phenomenon; b) as a parameter of a function or feature sensitive to some physical phenomenon extension. Scale is often invoked in a confused way by mixing these two meanings, as a formal element and as a natural quantity. In this article, the hidden scale parameter e referred exclusively to meaning b).

The probabilistic model presented in this work mixes several scale conditioned features which address different physical phenomena, i.e., different image characteristics. For example the physical quantity referred by the scale parameter e of our "Gabor and spreading" contrast feature $\mathbf{D}_s^e$ is the size of the structuring element while for the local entropy based feature H_s^e it is the width of the sampling window. A fundamental question arises on how these two quantities measuring different image charateristics are related. This is indeed a critical problem which has not been adressed explicitly by authors who developped multiscale detection methods. For instance, in [30] (Section II-A), the overall detector selects a scale using a multiscale LoG filter and feeds it to a Harris-like filter for the final output. Compared with these methods our approach tries to solve the scale parameter correspondence problem between different features by introducing a learning phase able to compensate for a

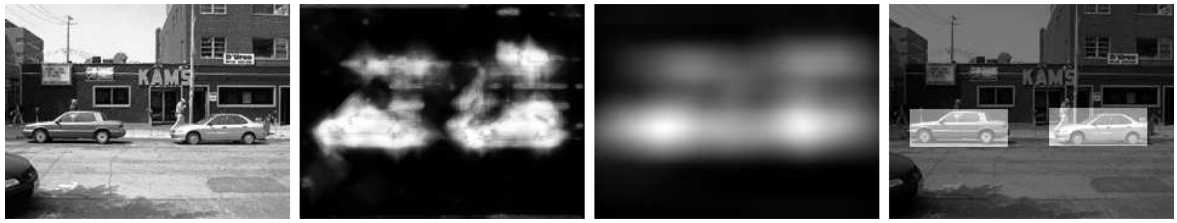


Fig. 13. Steps of saliency map postprocessing. From left to right: Original image, saliency map, filtered saliency map, and window centered at local maximum.

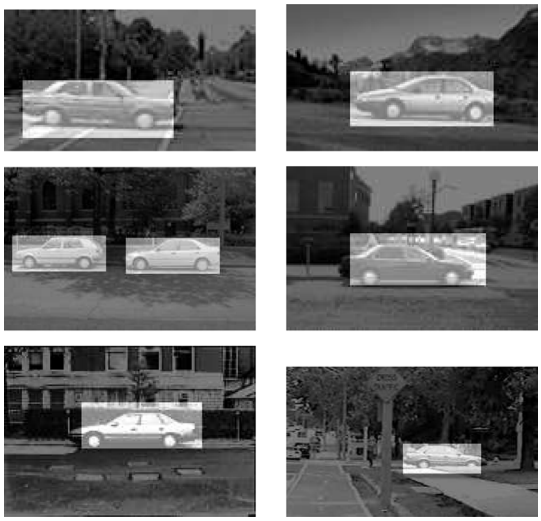


Fig. 14. Examples of good detections. Highlighted areas correspond to a window centered at a local maximum of the filtered saliency map.

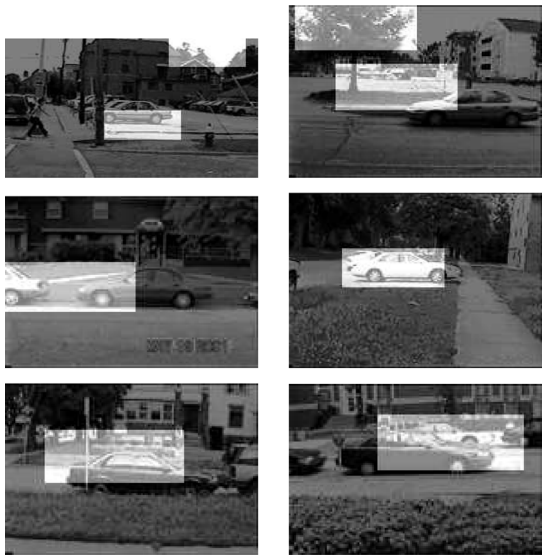


Fig. 15. Examples of detection candidates rejected by the acceptance criterion described in [28].

possible "bad correspondence" between the two feature scales. In case this correspondence is known (empirically or theoretically), one can easily incorporate it in the mixture model.

TABLE II
TEST SET I. REPARTITION OF DETECTIONS FOR VARIOUS THRESHOLDS AND GAUSSIAN SMOOTHING WITH STANDARD DEVIATION OF 15 PIXELS. VALIDATION CRITERION IS *FOCUS* (SEE TEXT FOR DEFINITION)

Threshold	**0.30**	**0.40**	**0.50**	**0.60**	**0.70**	**0.80**
nP - NP	0	3	4	7	12	31
NF	519	447	371	302	213	138
Precision	27.8%	30.6%	34.6%	39.0%	46.9%	55.0%
Recall	100.0%	98.5%	98.0%	96.5%	94.0%	84.5%

TABLE III
TEST SET II. REPARTITION OF DETECTIONS FOR VARIOUS THRESHOLDS AND GAUSSIAN SMOOTHING WITH STANDARD DEVIATION OF 15 PIXELS. VALIDATION CRITERION IS *FOCUS* (SEE (SEE TEXT FOR DEFINITION)

Threshold	**0.50**	**0.60**	**0.70**	**0.80**
nP - NP	0	3	6	10
NF	689	600	498	367
Precision	16.8%	18.5%	21.1%	26.0%
Recall	100.0%	97.8%	95.7%	92.8%

TABLE IV
TEST SET I. REPARTITION OF DETECTIONS FOR VARIOUS THRESHOLDS AND A SPATIAL AVERAGING WITH WINDOW OF SIZE 40×80. VALIDATION CRITERION IS *REGION* (SEE TEXT FOR DEFINITION)

Threshold	**0.20**	**0.30**	**0.40**	**0.50**	**0.60**	**0.70**
nP - NP	34	35	36	41	53	71
NF	181	138	110	87	57	26
Precision	47.8%	54.5%	59.9%	64.6%	72.1%	83.2%
Recall	83.0%	82.5%	82.0%	79.5%	73.5%	64.5%

TABLE V
TEST SET II. REPARTITION OF DETECTIONS FOR VARIOUS THRESHOLDS AND A MULTISCALE DETECTION. VALIDATION CRITERION IS *REGION* (SEE TEXT FOR DEFINITION)

Threshold	**0.40**	**0.70**	**1.00**	**1.20**	**1.40**	**1.50**
nP - NP	88	88	90	91	92	93
NF	554	399	284	237	177	158
Precision	8.4%	11.3%	14.7%	16.8%	21.0%	22.5%
Recall	36.7%	36.7%	35.3%	34.5%	33.8%	33.1%

TABLE VI
RESULTS PRESENTED IN [28]. VALIDATION CRITERION IS *REGION* (SEE TEXT FOR DEFINITION). IN THEIR ARTICLE, AGARWAL *et al.* ANNOUNCE TWO KINDS OF RESULTS BASED ON TWO DIFFERENT POSTPROCESSING

Test set	**I**	**I**	**II**	**II**
Post. proc	**1**	**2**	**1**	**2**
nP - NP	31	17	69	112
NF	140	557	215	1216
Precision	54.7%	24.7%	24.6%	8.4%
Recall	84.5%	91.5%	50.4%	80.6%

B. Object Detection

Most recent successful object detection approaches, model based, exploit very specific object features: This has the double advantage of limiting false detections, and providing an approximation of the object extension as the union of such characteristic features. They rely on a heavy learning phase requiring well annotated data bases sampling the true object distribution, which is a very strong operational requirement.

The approach described in this paper, more data driven, can be calibrated using few examples (five positives in the experiments described above). However, it requires that the objects of interest are "naturally" salient, i.e., stand with an idiosyncratic density of local contrasts described by a local scale. A shape boundary can be extracted from the saliency map, on the condition that all object parts have a minimal saliency level.

The tradeoff between model based and data driven approaches cannot be fixed easily for all contexts. Typically, in the car detection problem, a model relying on few but very selective localized features like tyres is likely to perform better than a generic contour based data description. This is one of the main motivation of patch based object models. In the plane detection problem, however, where any object part is likely to be confused with another element in the airport environment, finding localized discriminating features is harder. This poorly discriminating environment worsens if rotation invariance is required. Global object detection and shape boundary extraction in this context should rather rely on decision schemes exploiting the joint occurrence of several weakly selective features.

VII. CONCLUSION

This paper has presented an original approach for focusing on regions of interest in large images. It is based on a probabilistic model with few parameters that carries out pixels classification from two series of multiscale features, one bringing contrast information, the other bringing scale information. The model is able to handle objects of various sizes combining in an optimal way scale sensitive detectors. The multiscale features can be very simple, and their computation can be parallelized, allowing to process quickly the whole image. The probabilistic formalization gives rise to a learning scheme (EM algorithm) able to adapt the model parameters with few samples. Experiments on real aerial images and on urban scenes illustrate the algorithm behavior.

REFERENCES

[1] F. Fleuret and D. Geman, "Coarse-to-fine face detection," *Int. J. Comput. Vis.*, vol. 41, no. 1/2, pp. 85–107, 2001.

[2] Y. Amit and D. Geman, "A computational model for visual selection," *Neur. Comput.*, vol. 11, no. 7, pp. 1691–1715, 1999.

[3] M. C. Burl, M. Weber, and P. Perona, "A probabilistic approach to object recognition using local photometry and global geometry," *Lect. Notes Comput. Sci.*, vol. 1407, pp. 628–641, 1998.

[4] R. Fergus, P. Perona, and A. Zisserman, "Object class recognition by unsupervised scale-invariant learning," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2003, vol. II, pp. 264–271.

[5] D. Nair and J. Aggarwal, "Bayesian recognition of targets by parts in second generation forward looking infrared images," *Image Vis. Comput.*, no. 18, pp. 849–864, 2000.

[6] C. Olson and D. Huttenlocher, "Automatic target recognition by matching oriented edge pixels," *IEEE Trans. Image Process.*, vol. 6, no. 1, pp. 103–113, Jan. 1997.

[7] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2001, vol. 1, pp. 511–518.

[8] T. Zhao and R. Nevatia, "Car detection in low resolution aerial images," *Image Vis. Comput.*, vol. 21, no. 8, pp. 693–703, Aug. 2003.

[9] R. Stoica, X. Descombes, and J. Zerubia, "A Gibbs point process for road extraction in remotely sensed images," *Int. J. Comput. Vis.*, vol. 57, no. 2, pp. 121–136, 2004.

[10] C. Lin and R. Nevatia, "Building detection and description from a single intensity image," *Comput. Vis. Image Understand.*, vol. 72, no. 2, pp. 101–121, 1998.

[11] D. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, Nov. 2004.

[12] G. Dorko and C. Schmid, "Selection of scale-invariant parts for object class recognition," in *Proc. Int. Conf. Computer Vision*, 2003, pp. 634–640.

[13] B. Leibe and B. Schiele, "Scale invariant object categorization using a scale-adaptive mean-shift search," in *Proc. DAGM Annu. Pattern Recognition Symp.*, 2004, vol. 3175, pp. 145–153.

[14] L. Fei-Fei, R. Fergus, and P. Perona, "Learning generative visual models from few training examples: an incremental bayesian approach tested on 101 object categories," presented at the CVPR, Workshop on Generative-Model Based Vision 2004.

[15] F. Jurie and C. Schmid, "Scale-invariant shape features for recognition of object categories," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2004, vol. II, pp. 90–96.

[16] B. Leibe and B. Schiele, "Analyzing appearance and contour based methods for object categorization," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2003, vol. II, pp. 409–415.

[17] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 11, pp. 1254–1259, Nov. 1998.

[18] D. Walther, L. Itti, M. Riesenhuber, T. Poggio, and C. Koch, "Attentional selection for object recognition—A gentle way," in *Proc. Brit. Mach. Vis. Conf.*, 2002, pp. 472–479.

[19] Y. Amit and M. Mascaro, "An integrated network for invariant visual detection and recognition," *Image Vis. Comput.*, vol. 43, pp. 2073–2088, Sep. 2003.

[20] T. Kadir and M. Brady, "Saliency, scale and image description," *Int. J. Comput. Vis.*, vol. 45, no. 2, pp. 83–105, Nov. 2001.

[21] T. Lindeberg, "Feature detection with automatic scale selection," *Int. J. Comput. Vis.*, vol. 30, no. 2, pp. 79–116, Nov. 1998.

[22] B. Chalmond, C. Graffigne, M. Prenat, and M. Roux, "Contextual performance prediction for low-level image analysis algorithms," *IEEE Trans. Image Process.*, vol. 10, no. 7, pp. 1039–1046, Jul. 2001.

[23] D. Casasent and A. Ye, "Detection filters and algorithm fusion for atr," *IEEE Trans. Image Process.*, vol. 6, no. 1, pp. 114–125, Jan. 1997.

[24] A. Jain, N. Ratha, and S. Lakshmanan, "Object detection using gabor filters," *Pattern Recognit.*, vol. 30, no. 2, pp. 295–309, Feb. 1997.

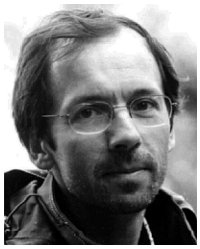
[25] H. Park and H. Yang, "Invariant object detection based on evidence accumulation and gabor features," *Pattern Recognit. Lett.*, vol. 22, no. 8, pp. 869–882, Jun. 2001.

[26] A. Agresti, *Categorical Data Analysis*, 2nd ed. New York: Wiley, 2002.

[27] B. Chalmond, *Modeling and Inverse Problems in Image Analysis*. New York: Springer-Verlag, 2003.

[28] S. Agarwal, A. Awan, and D. Roth, "Learning to detect objects in images via a sparse, part-based representation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 26, no. 11, pp. 1475–1490, Nov. 2004.

[29] B. Leibe and B. Schiele, "Interleaved object categorization and segmentation," presented at the Brit. Machine Vision Conf. 2003.
[30] K. Mikolajczyk and C. Schmid, "Scale and affine invariant interest point detectors," *Int. J. Comput. Vis.*, vol. 60, no. 1, pp. 63–86, 2004.



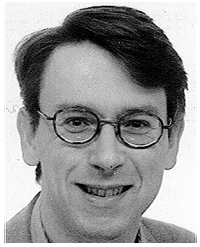
Bernard Chalmond received the Ph.D. degree in applied mathematics from the University of Paris-Sud, Orsay, France, in 1979, with a thesis on dealing with change detection models in times series.

After spending five years as Assistant Professor of medical informatics at the University Medical Center of Dijon, France, in 1984, he joined the Department of Applied Mathematics, University of Paris-Sud. Since 1991, he has been a Full Professor of information sciences with the Department of Physics, University of Cergy-Pontoise. He is affiliated with the Center of Applied Mathematics (CMLA), Ecole Normale Supérieure, Cachan.

Dr. Chalmond is performing research in the areas of data analysis and practical applications of computer vision in industry. He has been involved in the creation of companies in bioengineering (1985) and numerical imaging (1991). His main interests have covered statistical methods and algorithms for stochastic processes, spline approximation, multivariate data analysis, 3-D object reconstruction, and Monte Carlo simulation.

Benjamin Francesconi received the higher school engineering degree and the M.Sc. degree (optics, image, signal) from the Ecole Nationale Supérieure de Physique de Marseille, Marseille, France, in 2002. He is currently pursuing the Ph.D. degree at the Information Processing and Modeling Department (DTIM), ONERA, Châtillon, France.

His main research interest is detection and recognition of objects in aerial images.



Stéphane Herbin received the engineering degree from the Ecole Supérieure d'Electricité (Supélec), France, in 1990, the M.Sc. degree in electrical engineering from the University of Illinois at Urbana-Champaign, Urbana, in 1992, and the Ph.D. degree in applied mathematics from the Ecole Normale Supérieure de Cachan, Cachan, France, in 1997.

He was with the Aérospatiale Matra Missiles (now MBDA) from 1998 to 2000. Since 2000, he has been with the Information Processing and Modeling Department, ONERA. His main research interest is in stochastic modeling and analysis for object detection and recognition.