Statistical Context Priming for Object Detection

Antonio Torralba Pawan Sinha

Department of Brain and Cognitive Science, MIT
45 Carleton Street, E25-201, Cambridge, MA 02142
torralba@ai.mit.edu, sinha@ai.mit.edu
International Conference on Computer Vision, 2001

Abstract

There is general consensus that context can be a rich source of information about an object's identity, location and scale. However, the issue of how to formalize contextual influences is still largely open. Here we introduce a simple probabilistic framework for modeling the relationship between context and object properties. We represent global context information in terms of the spatial layout of spectral components. The resulting scheme serves as an effective procedure for context driven focus of attention and scale-selection on real-world scenes. Based on a simple holistic analysis of an image, the scheme is able to accurately predict object locations and sizes.

a)  b) 

Figure 1. a) Structured world: background and objects properties are correlated. b) Unstructured world: no rules constraint the possible arrangements. Notice how veridical context facilitates localization of the pedestrians in fig. (a) relative to fig. (b).

1 Introduction

In the real world, there exists a strong relationship between the environment and the objects that can be found within it. Experiments in scene perception [1] have shown that the human visual system makes extensive use of these relationships for facilitating object detection and recognition (Fig. 1: where are the pedestrians?). It seems that the visual system first processes context information in order to index object properties. From a computational point of view, this approach makes sense only if the context can be processed in a simple stage, simpler than the detection and the recognition of single objects. Context can play a useful role in object detection in at least two ways. First, it can facilitate object identification when the local intrinsic information about object structure is insufficient (say when the object appears at very small scales in an image). Second, even when objects can be identified via intrinsic information, context can simplify the object discrimination by cutting down on the number of object categories, scales and positions that need to be considered. Most current approaches to object detection are not designed to make use of contextual information. For instance, a pedestrian detection algorithm will be equally performing in both Fig. 1.a and Fig. 1.b. For most machine vision systems, no performance benefits accrue by the inclusion of veridical context cues.

One way of defining the 'context' of an object in a scene is in terms of other previously recognized objects within the scene. The drawback of this conceptualization is that it renders the complexity of context analysis to be at par with the problem of individual object recognition. An alternative view of context, which is algorithmically more attractive, relies on using the entire scene information holistically. This dispenses with the need for identifying individual objects within a scene. This is the viewpoint we shall adopt in the work presented here. Our goal is to develop a scheme for representing context information and to demonstrate its role in facilitating individual object detection. We shall show that context can 'prime' an object detection system by providing strong cues for location and scale selection. We show that the context processing stage is as simple as the recognition of an isolated object under controlled con-

ditions of location, size and pose. Therefore, context is an efficient shortcut for object detection and recognition even when, in principle, the task can be solved ignoring context.

2 Statistical object detection

In a probabilistic framework, the problem of object detection requires the evaluation of the function:

$$P(\vec{p}, \sigma, \vec{x}, o_n \mid \vec{v}) \tag{1}$$

This is the conditional probability density function (PDF) of the presence of an object o_n, at the spatial location $\vec{x}$, with pose $\vec{p}$ and size σ given a set of image measurements $\vec{v}$. $\vec{v}$ may be the pixel intensity values, the color distributions, the output of multiscale oriented band-pass filters, etc. Object detection and recognition requires the evaluation of this PDF at different locations in the parameter space defined by $(\vec{p}, \sigma, \vec{x}, o_n)$ (e.g. [9, 11, 13]).

2.1 Local features

Note that as written in (1), $\vec{v}$ refers to the image measurements at all spatial locations. Thus, $\vec{v}$ has a very high dimensionality. In order to reduce the complexity, it is assumed that the regions surrounding the object have independent features with respect to the object presence. Therefore, the PDF that is actually used by statistical approaches is [9, 11, 13]:

$$P(\vec{p}, \sigma, \vec{x}, o_n \mid \vec{v}) \simeq P_l(\vec{p}, \sigma, \vec{x}, o_n \mid \vec{v}_{B(\vec{x},\epsilon)}) \tag{2}$$

$\vec{v}_{B(\vec{x},\epsilon)}$ is a set of local image measurements in a neighborhood B of the location $\vec{x}$ with a size defined by $\epsilon = g(\vec{p}, \sigma)$ which is a function of the pose and size of the object. Eq. (2) formalizes the main principle underlying the classic approach for object detection: *the only image features that are relevant for the detection of an object at one spatial location are the features that potentially belong to the object and not to the background*. For instance, in a template-matching paradigm, the object detection is performed by the computation of similarities between image patches and a template built directly from the object. The image patches that do not satisfy the similarity criteria are discarded and modeled as noise with particular statistical properties.

2.2 Context features

In this paper, we shall formalize the intuition that there is a strong relationship between the background and the objects that can be found inside of it. The background can not only provide an estimate of the likelihood of finding an object (for example, one is unlikely to find a car in a room), it can also indicate the most likely position and scales at which an object might appear (e.g. pedestrians on walkways in an urban area). In order to model the context features, we split image measurements $\vec{v}$ in two sets:

$$\vec{v} = \left\{ \vec{v}_{B(\vec{x},\epsilon)},\ \vec{v}_{\overline{B}(\vec{x},\epsilon)} \right\} = \{\vec{v}_L,\ \vec{v}_C\} \tag{3}$$

where B refers to the local spatial neighborhood of the location $\vec{x}$ and $\overline{B}$ refers to the complementary spatial locations. Assuming that, given the presence of an object o_n at the location $\vec{x}$, the intrinsic object features and context features are independent, we can write:

$$P(\vec{v} \mid \vec{p}, \sigma, \vec{x}, o_n) = P_l(\vec{v}_L \mid \vec{p}, \sigma, \vec{x}, o_n) \cdot P_c(\vec{v}_C \mid \vec{p}, \sigma, \vec{x}, o_n)$$

The two conditional PDFs obtained refer to:

- *Local evidence*: $P_l(\vec{p}, \sigma, \vec{x}, o_n \mid \vec{v}_L)$. It is the object conditional PDF function given the set of local features $\vec{v}_L$. If the local measurements are appropriate, the PDF has strong and narrow maxima providing a high confidence. Its evaluation requires exhaustive spatial and multiscale search, resulting in a computationally expensive procedure.
- *Context priming*: $P_c(\vec{p}, \sigma, \vec{x}, o_n \mid \vec{v}_C)$. It is the object conditional PDF function given the set of context features $\vec{v}_C$. We do not expect this PDF to have strong, narrow maxima. Therefore, it provides *priors* on the object presence, location, scale and pose. By appropriately choosing a low dimensional context representation, the PDF can be computed efficiently.

From a computational point of view, context priming reduces the set of possible objects and therefore the number of features for discriminating between objects. It reduces the need for multiscale search and focuses computational resources into the more likely spatial locations. Therefore, we propose that the first stage of an efficient computational procedure for object detection comprises the evaluation of the PDF P_c.

3 Context priming

In this paper we will study the information available in the function P_c. We apply the Bayes rule successively in order to split the PDF P_c in four factors that model four kinds of context priming:

$$\begin{aligned} P_c(\vec{p}, \sigma, \vec{x}, o_n \mid \vec{v}_C) &= P_p(\vec{p} \mid \sigma, \vec{x}, o_n, \vec{v}_C) \\ P_s(\sigma \mid \vec{x}, o_n, \vec{v}_C) P_f(\vec{x} \mid o_n, \vec{v}_C) P_o(o_n \mid \vec{v}_C) \end{aligned} \tag{4}$$

The meanings of the four factors are, from right to left:

- *Object priming*: $P_o(o_n \mid \vec{v}_C)$. It selects the most likely objects given context information.

- *Focus of attention*: $P_f(\vec{x} \,|\, o_n, \vec{v}_C)$. It gives the most likely locations for the presence of object o_n given context information.
- *Scale selection*: $P_s(\sigma \,|\, \vec{x}, o_n, \vec{v}_C)$. It gives the most likely scales (sizes, distances) of the object o_n at different spatial locations given context information.
- *Pose and shape priming*: $P_p(\vec{p}, \,|\, \sigma, \vec{x}, o_n, \vec{v}_C)$. It gives the more likely (prototypical) shapes (point of views and poses) of the object o_n in the context $\vec{v}_C$.

Although these kinds of context priming have been shown to be important in human vision [1], computational models of object detection typically ignore the information available from the context.

4 Scene/context description

The definition of the context information given by eq. (3), $\vec{v}_C = \vec{v}_{\overline{B}(\vec{x},\epsilon)}$, has a very high dimensionality and depends on the pose, size and object location. In this section we show how context features can be represented in a low dimensional space without sacrificing relevant information.

4.1 Holistic representation

There are many examples of holistic representations in the field of object recognition. In contrast to parts-based schemes that detect and recognize objects based on an analysis of their constituent elements, holistic representations do not attempt to decompose an object into smaller entities. However, in the domain of scene recognition, most schemes have focused on 'parts-based' representations. Scenes are encoded in terms of their constituent objects and their mutual spatial relationships. But this requires the detection and recognition of objects as the first stage. Recent works have taken a different approach in which the scene is represented as a whole unity [10], as if it was an individual object, without splitting it into constituent parts (e.g. [5, 8, 10, 15, 16, 17, 18]). Previous studies have shown that the elements that seem to be relevant for discrimination between different scenes are: 1) The spatial structures (e.g. [5, 8, 10, 15, 16, 17, 18]): Different structural elements (e.g., buildings, road, tables, walls, with particular orientation patterns, smoothness/roughness) compose each context (e.g., rooms, streets, shopping center). 2) The spatial organization (e. g. [2, 8, 16, 17]): The structural elements have particular spatial arrangements. Each context imposes certain organization laws. 3) The color distribution [2, 5, 8, 15, 18]. As described below, we propose a low dimensional holistic representation that encodes the structural scene properties [10]. Color is not taken into account in this study, although the framework can be extended to include this attribute.

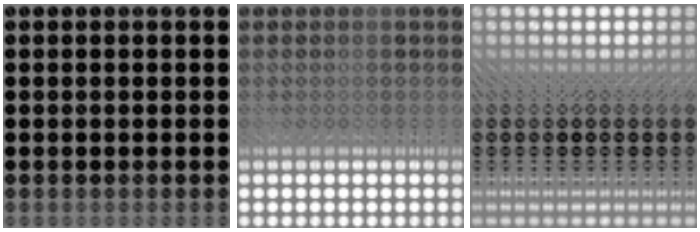

Figure 2. The first three PCs of the WFT at 16x16 locations and $r = 16$ pixels.

4.2 Spatial layout of main spectral components

We will use the magnitude of the Windowed Fourier transform (WFT) for describing the local structures of the scene. The WFT is defined as:

$$I(x, y, f_x, f_y) = \sum_{x',y'=0}^{N-1} i(x', y')\, h_r(x' - x, y' - y)\, e^{-j\,2\pi(f_x x' + f_y y')} \quad (5)$$

$i(x, y)$ is the input image and $h_r(x', y')$ is a hamming window with a circular support of radius r. A similar representation can be obtained by using multiscale oriented wavelet decomposition. We chose the WFT, as it can be easily visualized. In order to be tolerant to illumination variations and to reduce the sensitivity to spatial variations and contrast, we use the normalized local amplitude spectrum:

$$A(x, y, f_x, f_y) = \frac{|I(x, y, f_x, f_y)|}{I(x, y, 0, 0)\; std(x, y, f_x, f_y)} \quad (6)$$

with $std^2(x, y, f_x, f_y) = E\left[\left(A - \overline{A}\right)^2\right]$ where the expectation is approximated by averaging over the image database. This representation of the scene is of much higher dimensionality than $i(x, y)$ (in fact, due to the redundancy it is possible to invert the transformation for recovering the phase information discarded in eq. 6). To reduce the computations, we evaluated the local Fourier transforms only at 16x16 locations (with $r = 16$ pixels for the hamming window). In order to further reduce the dimensionality of the representation, we decomposed the local amplitude spectrum into its principal components (PC):

$$A(x, y, f_x, f_y) \simeq \sum_{n=1}^{N} a_n \psi_n(x, y, f_x, f_y) \quad (7)$$

with $a_n =< A(x, y, f_x, f_y), \psi_n(x, y, f_x, f_y) >$, where the functions ψ_n are the eigenfunctions of the covariance operator given by $A(x, y, f_x, f_y)$. Fig. 2 shows the first three PCs obtained. By using only a reduced set of PCs ($N = 60$

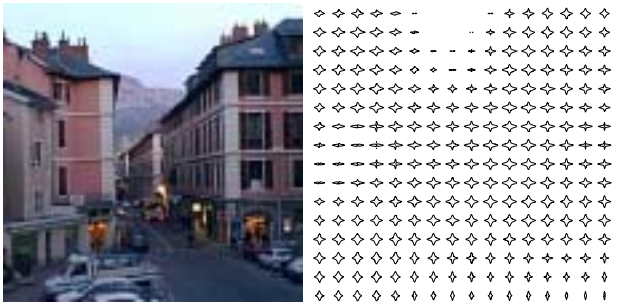

Figure 3. Spatial layout of main spectral components obtained from the first 20 PCs. The obtained layout captures the dominant orientations and scales at coarse image regions.

for the rest of the paper), eq. (7) provides an approximation of the layout of spectral components (see fig. 3). The coefficients $\{a_n\}_{n=1,N}$ encode the main spectral characteristics of the scene with a coarse description of their spatial arrangement. This provides the necessary degree of invariance with respect to objects arrangements, textures, and surfaces that are compatible with the same scene. In essence, $\{a_n\}$ is a holistic representation as all the regions of the image contribute to all the coefficients, and objects are not encoded individually [10]. In this representation the set of image measurements of eq. (1) is defined by $\vec{v} = A(x, y, f_x, f_y)$. We propose to use $\vec{v}_C = \{a_n\}_{n=1,N}$ as context features. This definition of context features differs from the one given in eq. (3) mainly because here $\vec{v}_C$ is computed from all the image measurements without discarding the ones belonging to the object o_n. When the object size is small with respect to the size of the image and $\vec{v}_C$ has a low dimensionality, the coefficients a_n are mostly determined by the context and not by the object. When the object is occupying a significant portion of the image, then $\vec{v}_C \sim \vec{v}_L$. We discuss the consequences of this fact later.

5 Context-driven scale selection and focus of attention

In order to illustrate the procedure, we focus on one object family: human heads in outdoor and indoor urban environments. Face detection is a very active field of research due to its many applications. The procedure can be extended to deal with other object families and other environmental categories.

5.1 PDF model and learning

As written in eq. (5), modeling context priming requires the estimation of conditional PDFs with the general form $P(\vec{u} \mid \vec{v})$. Here we adopt the mixture of gaussians model for the joint PDF:

$$P(\vec{u}, \vec{v}) = \sum_{i=1}^{M} b_i \, G(\vec{u}; \vec{u}_i, \mathbf{U}_i) G(\vec{v}; \vec{v}_i, \mathbf{V}_i) \tag{8}$$

with b_i being the weights of the local models, $\vec{u}$ being the output vector and $\vec{v}$ the input vector, and:

$$G(\vec{u}; \vec{u}_i, \mathbf{U}_i) = \frac{e^{-\frac{1}{2}(\vec{u}-\vec{u}_i)^T \mathbf{U}_i^{-1}(\vec{u}-\vec{u}_i)}}{(2\pi)^{L/2}|\mathbf{U}_i|^{1/2}} \tag{9}$$

L being the dimension of the vector $\vec{u}$. The parameters for the input distribution are $\vec{v}_i$ for the mean and $\mathbf{V}_i$ for the covariance matrix of the cluster i. The parameters for the output distribution are $\vec{u}_i$ and $\mathbf{U}_i$. The model includes a linear dependency between the mean of the output distribution and the input vector: $\vec{u}_i = \vec{a}_i + \mathbf{A}_i(\vec{v} - \vec{v}_i)$ [4]. The learning is performed by means of the EM algorithm (see [4] for a derivation of the learning equations). The database consists in 1700 pictures of 256^2 pixels. The scenes used spanned a range of categories and distances: indoors (rooms, restaurant, supermarket, stations, etc.) and outdoors (streets, shopping area, buildings, houses, etc.). For each picture the size and location of the heads was introduced by hand. The head heights ranged from 2 pixels to 250 pixels. For the learning stage we use one half of the database, and the other half is used for the testing stage. The results presented in the ensuing sections correspond to the mean performances averaged over several training trials.

5.2 Context-driven focus of attention

There are several studies modeling focus of attention. They are based on low-level saliency maps (without any high-level information relative to the task or context, e.g. [6, 7]) or they are only task driven (based on target models, e.g. [12, 9]). Common to all these models is the use of features in a *local-type* framework ignoring more high-level context information that is available in a *global-type* framework. In contrast to the cited approaches, the model we propose here is both task driven (looking for object o_n) and context driven (given global context information: $\vec{v}_C$). From an algorithmic point of view, focus of attention is important as it avoids expending computational resources in spatial locations with low probability of having the target. It also provides criteria for rejecting possible false detections that fall outside the primed region. When the target is small (a few pixels), the problem of detection using only local features is ill-posed (for instance, the first image in figure 4). In that case, some of the pedestrians are just scratches on the image. Similar scratches can be found in other locations of the picture. Due to context information, they are not considered as potential targets by the human visual system

Figure 4. Focus of attention based on global context configuration. Each of the four pairs shows the original image and the image multiplied by the function $p(\vec{x} \mid \vec{v}_C, heads)$ to illustrate the primed regions.

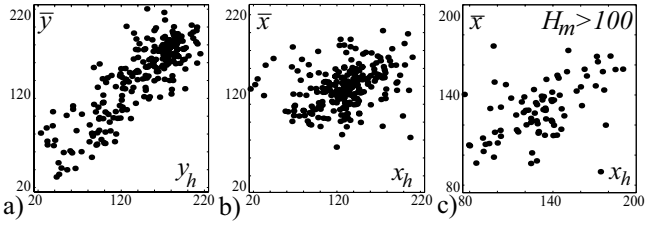

Figure 5. Estimation of heads locations. Coordinates $(x = 0, y = 0)$ corresponds to the left-upper image corner.

as they fall outside the 'pedestrian region' (see examples of this in figures 4, and 8). As mentioned in section 3, the problem of focus of attention can be formulated as the selection of the spatial locations that have the highest prior probability of containing the target object given context information. In our framework, it involves the evaluation of the PDF $P_f(\vec{x}|\, o_n, \vec{v}_C)$. The learning provides the relationship between the context and the more typical locations of the objects belonging to one family. For the PDF P_f we use the model given in eq. (8) which gives:

$$P_f(\vec{x}|\, o_n, \vec{v}_C) = \frac{\sum_{i=1}^{M} b_i\, G(\vec{x}; \vec{x}_i, \mathbf{X}_i) G(\vec{v}_C; \vec{v}_i, \mathbf{V}_i)}{\sum_{i=1}^{M} b_i\, G(\vec{v}_C; \vec{v}_i, \mathbf{V}_i)} \quad (10)$$

with $\vec{x}_i = \vec{a}_i + \mathbf{A}_i(\vec{v}_C - \vec{v}_i)$. The learning is realized as described in section 5.1. Very good results are obtained by modeling the PDF using only $M = 4$ clusters. Figure 4 shows several examples of images and the selected regions based on context features. From the PDF P_f we selected the region with $P_f(\vec{x}|\, o_n, \vec{v}_C) > th$ with th set experimentally in order to have, on average, a selected area of 33% of the size of the image. 87% of the heads present in the pictures of the testing set were inside the selected regions. We can estimate the center of the region of focus of attention as:

$$(\overline{x}, \overline{y}) = \int \vec{x} P_f(\vec{x}|o_n, \vec{v}_C) d\vec{x} = \frac{\sum_{i=1}^{M} b_i\, \vec{x}_i\, G(\vec{v}_C; \vec{v}_i, \mathbf{V}_i)}{\sum_{i=1}^{M} b_i\, G(\vec{v}_C; \vec{v}_i, \mathbf{V}_i)}$$

We can differentiate between two situations: A) for small object sizes, the context features $\vec{v}_C$ are not influenced by the presence and location of the target. In such situation, the primed region is determined only by the prior knowledge about the context and typical object locations (P_f). Fig. 5 compares the estimated $\overline{y}$ (fig. 5.a) and $\overline{x}$ (fig. 5.b) coordinates with the real coordinates of the center of the region that contains the heads in each picture (x_h, y_h). We can see that the y_h coordinate can be estimated quite well; the error $(y_h - \overline{y})$ has a gaussian distribution with zero mean and standard deviation of 26 pixels. However, the x_h coordinate is poorly estimated. The correlation coefficient between x_h and the obtained $\overline{x}$ is 0.35. The reason for these results is that the y_h coordinate is affected by the ground level, point of view and distance which are aspects that refer to properties of the context and observer (see section 6). However, context introduces little constraint on the coordinate x_h as, in general, people can have any x location in the scene (see fig. 4). B) For big object sizes (> 100 pixels vs. 256 for the image) the situation is different as the global features $\vec{v}_C$ are affected by the local features that belong the object. In such a case, $\vec{v}_C$ is affected by the exact location that the object has in the scene. Fig. 5.c shows the estimated $\overline{x}$ coordinate with respect to the x_h location of the heads in the picture. The correlation coefficient between x_h and $\overline{x}$ is 0.7.

5.3 Context-driven scale selection

Scale selection is a fundamental problem in computational vision. If scale information could be estimated by a low cost pre-processing stage, then subsequent stages of object detection and recognition would be greatly simplified by focusing the processing onto the only diagnostic/relevant scales. With that aim, Lindeberg [7] proposed a method for scale selection for the detection of low-level features such as edges, junctions, ridges and blobs when no a priori information about the nature of the picture is available. Here we show that prior knowledge about context provides a strong cue for scale selection for the detection of high-level structures as objects. The context in which the object is located restricts its possible locations and distances. In the model we propose, the preferred scale for a context $\vec{v}_C$ is given by the PDF $P_s(\sigma \mid \vec{x}, o_n, \vec{v}_C)$. For simplicity, we assume that the scale is independent of position: $P_s(\sigma \mid \vec{x}, o_n, \vec{v}_C) \simeq P_s(\sigma \mid o_n, \vec{v}_C)$. The model for the con-

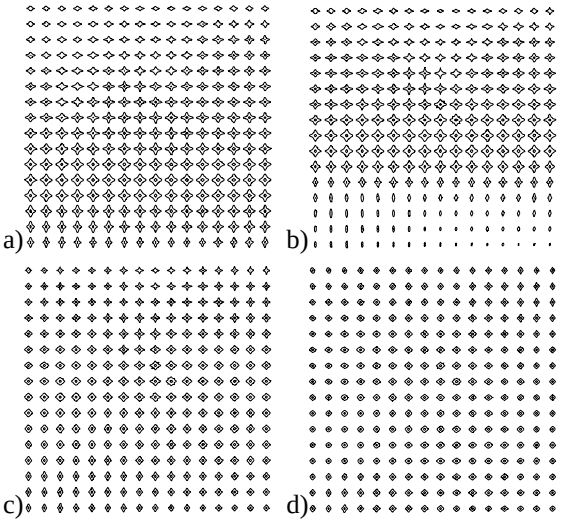

Figure 6. Spatial layout of spectral components corresponding to the four clusters obtained during the learning of the scale selection model.

ditional PDF is:

$$P_s(\sigma|\, o_n, \vec{v}_C) = \frac{\sum_{i=1}^{M} b_i\, G(\sigma; \sigma_i, \mathbf{S}_i) G(\vec{v}_C; \vec{v}_i, \mathbf{V}_i)}{\sum_{i=1}^{M} b_i\, G(\vec{v}_C; \vec{v}_i, \mathbf{V}_i)} \quad (11)$$

with $\sigma_i = a_i + \vec{A}_i^T(\vec{v}_C - \vec{v}_i)$. The model parameters are $(b_i, a_i, \vec{A}_i, \mathbf{S}_i, \vec{v}_i, \mathbf{V}_i)$ which are obtained after a learning stage. As here $o_n = heads$, we estimated σ as being the mean height H_m of the heads present in the picture (in logarithmic units): $\sigma = \log(H_m)$, with H_m given in pixels. Head height, which refers to the vertical dimension of the head, is mostly independent of head pose (variations in pose are mostly due to horizontal rotations). The preferred scale ($\overline{\sigma}$) given context information ($\vec{v}_C$) is estimated as the conditional expectation:

$$\overline{\sigma} = \int \sigma P_s(\sigma|\, o_n, \vec{v}_C) d\sigma = \frac{\sum_{i=1}^{M} \sigma_i\, b_i G(\vec{v}_C; \vec{v}_i, \mathbf{V}_i)}{\sum_{i=1}^{M} b_i\, G(\vec{v}_C; \vec{v}_i, \mathbf{V}_i)}$$

It is also possible to obtain a measure of the variance of head height for a given context in order to have a measure of the reliability of the estimation:

$$\overline{E^2} = \int (\sigma - \overline{\sigma})^2 P(\sigma|\, o_n, \vec{v}_C) d\sigma = \frac{\sum_{i=1}^{M} \mathbf{S}_i\, b_i\, G(\vec{v}_C; \vec{v}_i, \mathbf{V}_i)}{\sum_{i=1}^{M} b_i\, G(\vec{v}_C; \vec{v}_i, \mathbf{V}_i)}$$

The model reaches maximal performance with as few as $M = 4$ clusters. Fig. 6 shows the spectral layout associated with the input distribution for the four clusters obtained.

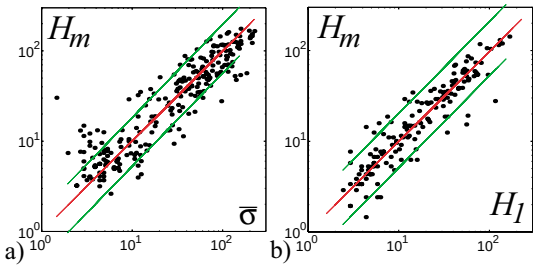

Figure 7. a) Results for context driven scale selection and b) Scale selection given the size of one of the objects.

The output distribution of the four clusters is centered at 5 (a), 11 (b), 32 (c) and 91 (d) pixels for the head height. The results for the estimation of H_m are given in figs. 7.a and 8. For 84% of the images, the estimated head height ($\overline{\sigma}$) was in the interval $\overline{\sigma} \in (H_m/2, H_m * 2)$, with H_m being the actual mean height of the heads in the picture and for 41% of the images $\overline{\sigma} \in (H_m/1.25, H_m * 1.25)$. Fig. 9 shows some of the images for which our scheme produced a wrong estimate. For comparison and in order to have an estimation of the best performances that can be attained from context-based information, we studied how good is the height of one head H_1 (selected at random among the heads present in the scene) as an estimator of the mean height H_m of the others heads in an image. For 90% of the pictures $H_1 \in (H_m/2, H_m * 2)$ (see Fig. 7.b).

6 Context properties

In general, the introduction of other object families into the model does not require learning new PDFs. As we show here, the PDFs of focus of attention and scale selection can be written as functions of a few context properties.

6.1 Absolute mean depth of the scene

The relative size (σ) of an object inside an image, depends on both the relative size σ_1 of the object at one fixed distance (e.g. 1 meter) and the actual distance D between the observer and the object: $\sigma = \sigma_1 - D$ (in logarithmic units and for linear measurements). If we can detect one object by applying a multiscale search, then, given σ_1, we can estimate the absolute distance at which it is located. This approach is the traditional approach for object detection and estimation of absolute depth from familiar object sizes. The approach we propose is quite different. As shown in section 5.3, using a holistic image representation we can infer the expected sizes for particular objects. This procedure implies having some knowledge about the absolute mean

Figure 8. Focus of attention and scale selection from global context information.

depth of the scene. In the scale priming model we can make explicit the mean depth of the scene (D) by writing:

$$P_s(\sigma \mid o_n, \vec{v}_C) \simeq \int P(\sigma \mid D, o_n) P_D(D \mid \vec{v}_C) dD \qquad (12)$$

The approximation comes from two assumptions: 1) Once the mean depth is specified, the object size is independent of context features: $P(\sigma \mid D, o_n, \vec{v}_C) = P(\sigma \mid D, o_n)$ and 2) the mean depth of the scene is independent of the object class that we want to detect: $P(D \mid o_n, \vec{v}_C) = P_D(D \mid \vec{v}_C)$. The function $P(\sigma \mid D, o_n)$ does not require a learning stage. A simple model of this PDF will consist of a gaussian distribution: $P(\sigma \mid D, o_n) \sim e^{-(\sigma - \sigma_1 - D)^2/\beta^2}$, with σ, σ_1 and D in logarithmic units. β^2 is the variance of the distribution and includes variability due to pose. Only the function $P_D(D \mid \vec{v}_C)$, which does not depend on o_n, requires a learning stage with a database of examples. If β^2 is small and $P_D(D \mid \vec{v}_C)$ is smooth then eq. (12) can be approximated by: $P_s(\sigma \mid o_n, \vec{v}_C) \simeq P_D(\sigma_1 - \sigma \mid \vec{v}_C)$. Then $\overline{\sigma} = \sigma_1 - \overline{D}$ with $\overline{D} = \int D P_D(D \mid \vec{v}_C) dD$. It should be noted that $P_D(D \mid \vec{v}_C)$ provides absolute depth information based only on monocular cues [17]. In such an approach, depth is provided by familiar context instead of familiar objects (Fig. 8 and 10 show pictures sorted according to absolute depth).

6.2 Horizon line

In a similar vein, when looking for an object, the focus of attention depends on few context properties. The $\overline{y}$ coordinate of the center of the focus of attention can be approximated by $\overline{y} \simeq (H - (a-h)/D)/(H(a-h)/D + 1)$, where H is the position of the horizon in the image, D is the mean depth of the scene, h is the height of the observer and a is the elevation of the object o_n with respect to ground level. In the case of $o_n = heads$, $a \sim h$ and we can approximate $\overline{y} \simeq H$. Fig. 10.b shows scenes (with similar mean depths) sorted with respect to the horizon level estimated using the $\overline{y}$ coordinate of the focus of attention.

Figure 9. Examples of errors chosen among the 10 biggest errors in scale priming.

To summarize, some scene properties such as absolute mean depth or ground level, that are usually believed to require additional sources of information (e.g. binocular vision) and local analysis (perspective lines), can be estimated by indexing them (or recognizing them) using a global context description.

7 Context familiarity

Scale selection and focus of attention depend on prior knowledge provided during the learning stage. One would expect the system to exhibit poor performance when analyzing context categories not included in the training. Therefore, it is important to have a measure of the familiarity in order to determine the reliability of the inferences. Familiarity of a context is quantified by the probability that it belongs to the set of context categories Ω used in the learning, given a set of context features $\vec{v}_c$. That is the conditional PDF $P(\Omega|\vec{v}_C)$, with: $P(\Omega|\vec{v}_C) = P(\vec{v}_C|\Omega)P(\Omega)/P(\vec{v}_C)$. This requires the evaluation of the distribution of the context features in the learning set Ω, and the total distribution of context features given any context category which can be written: $P(\vec{v}_C) = P(\vec{v}_C|\Omega)P(\Omega) + P(\vec{v}_C|\overline{\Omega})P(\overline{\Omega})$. The in class PDFs $P(\vec{v}_C|\Omega)$ and out of class PDF $P(\vec{v}_C|\overline{\Omega})$ are

a)
b)

Figure 10. Pictures sorted according to: a) Mean depth, and b) ground level.

modeled by a single gaussian. We assume $P(\Omega) = P(\overline{\Omega}) = 0.5$. One scene is considered familiar if $P(\vec{v}_C|\Omega) > P(\vec{v}_C|\overline{\Omega})$. Several tests using different combinations of familiar contexts and unfamiliar contexts (urban vs. portraits, urban vs. natural, etc.) yield results above 85% of correct rejection of unfamiliar context with less that 15% of rejection of familiar scenes.

8 Conclusion

We have shown that object locations and scales can be inferred from a simple holistic representation of context, based on the spatial layout of spectral components. The results lend credence to the intuition that the integration of contextual analysis into computational models for object detection should yield more efficient systems capable of making use of regularities in real-world scenes. However, several interesting issues remain open. These include the study of other context representations, the integration of the model in a comprehensive system for object detection and comparing the model's performance with that of human subjects on object localization tasks in large scenes. The last enterprise will likely suggest ways in which our model can be further refined.

References

[1] Biederman, I., Mezzanotte, R.J., and Rabinowitz, J.C. 1982. Scene perception: detecting and judging objects undergoing relational violations. *Cognitive Psychology*, 14:143–177.

[2] Carson, C., Belongie, S., Greenspan, H., and Malik, J. 1997. Region-based image querying. *Proc. IEEE W. on Content-Based Access of Image and Video Libraries*, pp: 42–49.

[3] De Bonet, J. S., and Viola, P. 1997. Structure driven image database retrieval. *Adv. in Neural Information Processing* **10**.

[4] Gershnfeld, N. *The nature of mathematical modeling*. Cambridge university press, 1999.

[5] Gorkani, M. M., and Picard, R. W. 1994. Texture orientation for sorting photos "at a glance". *Proc. Int. Conf. Pat. Rec.*, Jerusalem, Vol. I, 459–464.

[6] Itti, L., Koch, C., and Niebur, E. 1998. A model of saliency-based visual attention for rapid scene analysis. *IEEE Trans. Pattern Analysis and Machine Vision*, 20(11):1254–1259.

[7] T. Lindeberg. 1993. Detecting salient blob-like image structures and their scales with a scale-space primal sketch: A method for focus-of-attention. *International Journal of Computer Vision*, 11(3):283–318.

[8] Lipson, P., Grimson, E., and Sinha, P. 1997. Configuration based scene classification and image indexing. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. Puerto Rico, pp 1007-1013.

[9] Moghaddam, B., and Pentland, A. 1997. Probabilistic Visual Learning for Object Representation. *IEEE Trans. Pattern Analysis and Machine Vision*, 19(7):696–710.

[10] Oliva, A., and Torralba, A. 2001. Modeling the shape of the scene: a holistic representation of the spatial Envelope. *International Journal of Computer Vision*. 42(3):145–175.

[11] Papageorgiou, C., and Poggio, T. 2000. A trainable system for object detection. *International Journal of Computer Vision*, 38(1):15–33.

[12] Rao, R.P.N., Zelinsky, G.J., Hayhoe, M.M., and Ballard, D.H. 1996. Modeling saccadic targeting in visual search. NIPS'95. MIT press.

[13] Schiele, B., and Crowley, J.L. 1997. Recognition without Correspondence using Multidimensional Receptive Field Histograms. M.I.T. Media Laboratory, Perceptual Computing Section Technical Report No. 453

[14] Sirovich, L., and Kirby, M. 1987. Low-dimensional procedure for the characterization of human faces. *Journal of Optical Society of America*, 4, 519-524

[15] Szummer, M., and Picard, R. W. Indoor-outdoor image classification. In *IEEE intl. workshop on Content-based Access of Image and Video Databases*, 1998.

[16] Torralba, A., and Oliva, A. 1999. Scene organization using discriminant structural templates. *Proc. Of Int. Conf in Comp. Vision*, ICCV99, 1253-1258.

[17] Torralba, A., and Oliva, A. Depth perception from familiar structure. submitted.

[18] Vailaya, A., Jain, A., and Zhang, H. J. 1998. On image classification: city images vs. landscapes. *Pattern Recognition*, 31:1921–1935