massachusetts institute of technology — artificial intelligence laboratory

Contextual Priming for Object Detection

Antonio Torralba and Pawan Sinha

AI Memo 2001-020 September 2001
CBCL Memo 205

© 2001 massachusetts institute of technology, cambridge, ma 02139 usa — www.ai.mit.edu

Abstract

There is general consensus that context can be a rich source of information about an object's identity, location and scale. In fact, the structure of many real-world scenes is governed by strong configurational rules akin to those that apply to a single object. Here we introduce a simple probabilistic framework for modeling the relationship between context and object properties based on the correlation between the statistics of low-level features across the entire scene and the objects that it contains. The resulting scheme serves as an effective procedure for object priming, context driven focus of attention and automatic scale-selection on real-world scenes.

Keywords

Context, image statistics, object recognition, focus of attention, automatic scale selection, object priming

The authors would especially like to thank A. Oliva and W. Richards for fruitful discussions. The authors also want to thanks D. Kersten for comments about this work. This research was supported in part by funds from the Alfred P. Sloan Fellowship for neuroscience and the Defense Advanced Research Projects Agency.

I. Introduction

In the real world, there exists a strong relationship between the environment and the objects that can be found within it. There is increasing evidence of an early use of contextual information in human perception. Experiments in scene perception and visual search (Palmer, 1975; Biederman et al, 1982; De Graef et al, 1990; Henderson and Hollingworth, 1999; Chun and Jiang, 1998) have shown that the human visual system makes extensive use of these relationships for facilitating object detection and recognition suggesting that the visual system first processes context information in order to index object properties. In particular, scene recognition experiments show that information about scene identity maybe available before performing a more detailed analysis of the individual object (Potter, 1975; Biederman, 1988; Schyns and Oliva, 1994; Oliva and Schyns, 1997; Rensink et al., 1997). However, object-centered approaches dominate the research in computational vision. Object-centered representations use exclusively intrinsic object features for performing object detection and recognition tasks (e.g. Papageorgiou and Poggio, 2000; Schiele and Crowley, 2000; Rao et al, 1996; Moghaddam and Pentland, 1997).

The structure of many real-world scenes is governed by strong configurational rules akin to those that apply to a single object. In such situations, contextual information can provide more relevant information for the recognition of an object than the intrinsic object information (fig. 1). One way of defining the 'context' of an object in a scene is in terms of other previously recognized objects within the scene. However, in such a framework, the context representation is still object-centered as it requires object recognition as a first step. As shown in (Oliva and Torralba, 2001) it is possible to build a representation of the scene that bypasses object identities, in which the scene is represented as a single entity, holistically. The representation is based on the differential regularities of the second order statistics of natural images when considering different environments. Our goal here is to use such a scheme for including context information in object representations and to demonstrate its role in facilitating individual object detection (Torralba and Sinha, 2001a). The approach is based on using the differences of the statistics of low-level features in real-world images when conditioning those statistics to the presence/absence of objects and their locations and sizes.

The paper is organized as follows: in section 2 we review the role of context and discuss some of the past work on context-based object recognition. Section 3 formalizes the statistical framework used in this work for including context information in the object detection task. Section 4 details the contextual representation. Section 5 describes the image database used for our experiments. Sections 6, 7 and 8 describe respectively object priming, context-driven focus of attention and automatic context-driven scale selection.

II. Context

A. The role of context

Under favorable conditions, the multiplicity of cues (color, shape, texture) in the retinal image produced by an object provides enough information to unambiguously determine the object category. Under such high quality viewing conditions, the object recognition mechanisms could rely exclusively on intrinsic object features ignoring the background. Object recognition based on intrinsic object features can robustly handle many transformations such as displacement, rotation, scaling, changes in illumination, etc. Therefore, at least in principle, contributions from context do not seem necessary for object recognition.

However, in situations with poor viewing quality (caused, for instance, by large distances, or short acquisition times) context appears to play a major role in enhancing the reliability of recognition. This is because, in such circumstances, the analysis of intrinsic object information alone cannot yield reliable results (fig. 1.a). However, when the object is immersed in its typical environment, recognition of the object is reliable (figure 1.b). Under degradation, purely object-centered representations are not enough for accounting for the reliable object recognition performance of observers. In real-world scenes,

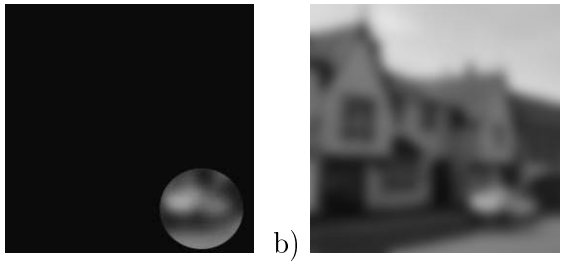

Fig. 1. The structure of many real-world scenes is governed by strong configurational rules akin to those that apply to a single object. In such situations, individual objects can be recognized even when the intrinsic information is impoverished, for instance due to blurring as shown above. In presence of image degradation (due to distance or fast scene scanning), object recognition mechanisms have to include contextual information in order to make reliable inferences. Recognition based on intrinsic features provides very poor performances when asking to observers. However, when the object is immersed in its typical environment, subjects experience a vivid recognition of the object.

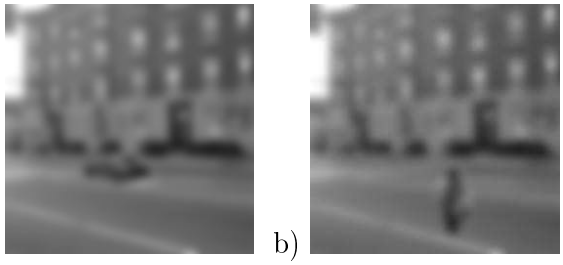

Fig. 2. In presence of image degradation (e.g. blur), object recognition is strongly influenced by contextual information. Recognition makes assumptions regarding objects identities based on its size and location in the scene. In this picture subjects describe the scenes as: a) a car in the street, and b) a pedestrian in the street. However, the pedestrian is in fact the same shape as the car except for a 90 degrees rotation. The non-typicality of this orientation for a car within the context defined by the street scene makes the car be perceived as a pedestrian. Without degradation, subjects can correctly recognize the rotated car due to the sufficiency of local features.

intrinsic object information is often degraded due to occlusions, illumination, shadows, peripheral vision and distance, leading to poor resolution and/or contrast. Therefore, the inclusion of context is mandatory in order to build efficient and reliable algorithms for object recognition. In the absence of enough local evidence about an object's identity, the scene structure and prior knowledge of world regularities provide the only information for recognizing and localizing the object (see figure 2).

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.

B. Context-based vision systems

Although computational vision is strongly dominated by object-centered approaches, there are previous studies that make use of contextual information for object recognition.

Hanson and Riseman (1978) proposed a system called VISIONS consisting of a collection of experts specialized for recognizing different types of objects. Each expert generates hypotheses about the presence and location of objects in the image. Based on hand-coded *if-then* rules, the system analyzes the consistency among the different hypotheses in order to arrive at reliable decisions. Built on a similar philosophy, the CONDOR system (Strat and Fischler, 1991) uses contextual information for object recognition. The system is based on a large number of hand-written rules that constitute the knowledge database of the system. A collection of rules (context sets) defines the conditions under which it is appropriate to use an operator to identify a candidate region or object. The candidates are then the inputs for other rules that will activate other vision routines. The ideal output of the system

is a labeled 3D model of the scene.

In (Fu, Hammond and Swain, 1994) context consists of prior knowledge about regularities of a reduced world in which the system has to operate. The regularities are fixed priors on the locations of the objects and the structure of the environment. In (Bobick and Pinhanez, 1995) context is used to validate and select vision routines. Context consists of a model (hand-written rules) of a reduced world in which the vision system has to work. The model includes a description of the 3D geometry of the scene in which the system works and also information about the camera's field of view. In (Moore et al, 1999), a prior model of a particular fixed scene and the identification of human motion constitute the context used for the recognition of objects.

Common to all these approaches are the use of an object-based representation of context and a rule-based expert system. The context is defined as a collection of objects or regions (already recognized or at least given candidate object labels). Predefined rules about the world in which the system is expected to operate produce reliable inferences using the candidates as input.

Other approaches use a statistical approach in order to model the joint distribution of N objects $O_1, ..., O_N$ within a scene (Haralick, 1983; Kitller, 2000; Song et al, 2000) given a set of local measurements $\vec{v}_1, ..., \vec{v}_N$ corresponding to each object:

$$P(O_1, ..., O_N, \vec{v}_1, ..., \vec{v}_N) \simeq \left[\prod_i^N P(\vec{v}_i|O_i)\right] P(O_1, ..., O_N) \tag{1}$$

The joint PDF is approximated then by assuming conditional independence between local image measurements. Therefore, contextual information is incorporated in the joint PDF by means of the factor $P(O_1, ..., O_N)$ which provides the prior probability of the different combinations of objects in the world. This formulation of context is object-centered and there is no attempt at identifying the context prior to the object recognition process.

As suggested in (Torralba and Oliva, 1999; Oliva and Torralba, 2001; Torralba and Sinha, 2001) the scene/context can be considered a single entity that can be recognized by means of a *scene-centered representation* bypassing the identification of the constituent objects. This is the approach that we adopt in the work presented here for representing contextual information. In the next section we introduce the main formalism of our approach.

III. Statistical object detection

In this section we introduce the general statistical framework on which the rest of the paper is based. Although other schemes can be used, a probabilistic framework yields a simple formulation of contextual influences on object recognition.

A. Local features and object-centered object likelihood

In the classical probabilistic framework, the problem of object detection given a set of image measurements $\vec{v}$ requires the evaluation of the object likelihood function:

$$P(O_n \,|\, \vec{v}) = \frac{P(\vec{v} \,|\, O_n)}{P(\vec{v})} P(O_n) \tag{2}$$

The function $P(O_n \,|\, \vec{v})$ is the conditional probability density function (PDF) of the presence of the object O_n given a set of image measurements $\vec{v}$. $\vec{v}$ may be the pixel intensity values, color distributions, output of multiscale oriented band-pass filters, shape features, etc. The notation O_n is used to summarize the properties of the object n: $O_n = \{o_n, \vec{x}, \sigma, ...\}$. Where o_n is the label of the object category (car, person, ...), $\vec{x}$ is the location of the object in image coordinates, and σ is the size of the object. This is not an exhaustive list and other parameters can be used to describe object properties like pose, illumination, etc.

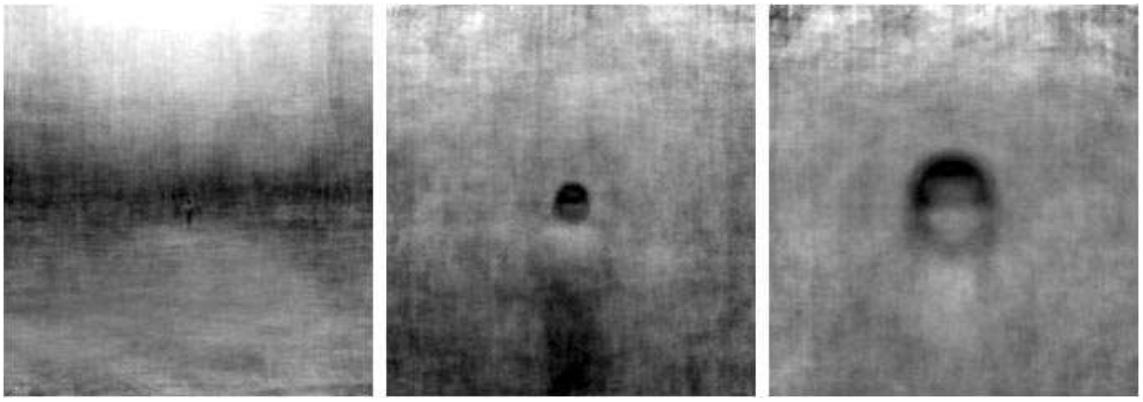

Fig. 3. Average of pictures containing heads in context at three different scales. Notice that the background does not average out to a homogeneous field but preserves some structure when centering the images with respect to the object.

Note that as written in eq. (2), $\vec{v}$ refers to the image measurements at all spatial locations. Although eq. (2) is the ideal PDF that has to be evaluated in order to detect an object, as all information provided by the image is taken into account, the high dimensionality of $\vec{v}$ makes the modeling of this PDF extremely ill-posed. In order to reduce the complexity, most object recognition schemes assume 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 for object recognition is (Moghaddam and Pentland, 1997; Schiele and Crowley, 2000):

$$P(O_n \,|\, \vec{v}) \simeq P(O_n \,|\, \vec{v}_L) = \frac{P(\vec{v}_L \,|\, O_n)}{P(\vec{v}_L)} P(O_n) \tag{3}$$

This function is the object-centered object likelihood. $\vec{v}_L = \vec{v}_{B(\vec{x},\epsilon)}$ is a set of local image measurements in a neighborhood B of the location $\vec{x}$ with a size defined by $\epsilon = g(\sigma)$ which is a function of the size of the object. The feature vector $\vec{v}_L$ is expected to have a low dimensionality. Eq. (3) formalizes the main principle underlying the classic approach for object detection using statistical paradigms or others: *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.* This assumption is common both to holistic approaches for object recognition and for part-based ones. For instance, in a holistic template-matching paradigm, the object detection stage is performed by the computation of similarities between image patches with different sizes and locations 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. The template used for the matching corresponds to an object-based view-centered representation. The context is thus treated as a collection of distractors but not as an entity that also conveys information about the object identity.

Such object-based approaches suffer from two important drawbacks: First, they cease to be effective when the image is so degraded (due to noise or lose of resolution due to viewing distances) that the intrinsic object information is insufficient for reliable recognition. Second, it requires exhaustive exploration of a large search space corresponding to different object models, locations and scales.

Instead of considering the background as a set of potential distractors that can produce false alarms for an object detection procedure, we propose to use the background statistics as an indicator of object presence and properties.

B. Context features and context-centered object likelihood

Fig. 3 shows a superposition of about 300 pictures containing people at three scales. The pictures have been corrected in translation so that the head is always centered in the image. However, there

is no correction in scale. Only the pictures that contain people with a given scale are averaged. The resulting images are composed of two parts: the first part is the average of the target object which in this case can still be easily recognized and segmented from the background due to the strong regular shape shared across different images and views. The second part is the background that does not average to a mean gray showing that there is also a regular pattern in the background structure when centering the image with respect to the object. More interestingly, the mean background has different properties when the object is at different scales.

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:

$$\vec{v} = \left\{\vec{v}_{B(\vec{x},\epsilon)},\, \vec{v}_{\overline{B}(\vec{x},\epsilon)}\right\} = \{\vec{v}_L,\, \vec{v}_C\} \tag{4}$$

where B refers to the local spatial neighborhood of the location $\vec{x}$ and $\overline{B}$ refers to the complementary spatial locations. The object likelihood function can be written as:

$$P(O_n \,|\, \vec{v}) = P(O_n \,|\, \vec{v}_L,\, \vec{v}_C) \tag{5}$$

Note that current object-centered computational approaches assume that $P(O_n \,|\, \vec{v}_L,\, \vec{v}_C) = P(O_n \,|\, \vec{v}_L)$ yielding eq. (3). In order to formalize contextual influences, we use Bayes' rule to write the object likelihood of eq. (5) as:

$$P(O_n \,|\, \vec{v}) = \frac{P(O_n, \vec{v})}{P(\vec{v})} = \frac{P(\vec{v}_L \,|\, O_n,\, \vec{v}_C)}{P(\vec{v}_L \,|\, \vec{v}_C)} P(O_n \,|\, \vec{v}_C) \tag{6}$$

The object likelihood function is decomposed in two factors: the first factor is the posterior distribution of local features when the object O_n is present in the context represented by $\vec{v}_C$. The normalization factor is the distribution of local features at the location $\vec{x}$ in the context $\vec{v}_C$. Note that this ratio differs from that of eq. (3) in that now all the probabilities are conditional with respect to contextual information. The inclusion of the context in the likelihood function might account for different appearances of the object as a function of the context (for instance, due to global illumination factors or the pose of the object). Although this point is of great interest and should be studied in more detail, in this paper we will focus on the study of the second factor which has received much less attention in computational vision and object recognition approaches. The second factor, the PDF $P(O_n \,|\, \vec{v}_C)$, provides context-based priors on object class, location and scale and it is of capital importance for insuring reliable inferences in situations where the image measurements produce ambiguous interpretations (see Jepson et al, 1996, for a discussion on this topic).

As it is written in eq. (5) the two set of features $\vec{v}_L$ (object-centered) and $\vec{v}_C$ (context-centered) seem to have a symmetric behavior. However, there is one important difference that breaks the symmetry. The definition of $\vec{v}_L$ (eq. 4) depends strongly on the object properties $O_n = (\sigma, \vec{x}, o_n, ...)$. In fact, we can write $\vec{v}_L(O_n)$. Therefore, the apparently simple object-centered likelihood is in fact $P_{O_n}(O_n|\vec{v}_L(O_n))$. That is, the local features that might be related to the object presence are a function of the object properties (for instance size and location). Moreover the likelihood function of the presence of an object, given local features, might also be a function of the object properties (for instance, accounting for variations in the pose of the object will require different parameters of the model of the PDF). This is the case of object detection paradigms that require exhaustive multiscale and multilocation search. This is not the case for contextual features $\vec{v}_C$. For small object sizes with respect to the background, the context features are only slightly affected by the object properties. The only problem is the high dimensionality of the context features vector. However, we will see in section IV that the

dimensionality can be drastically reduced by simple techniques such as PCA while preserving most of the information relevant for object priming. Therefore, we will show that context-centered object likelihood can be computed efficiently in a single step. The difference in the complexities of local and contextual features justifies the choice of the form of equation (6) in which contextual features are used to simplify the distribution of the local features.

C. Context-centered object likelihood

As many studies have already focused on the function $P(O_n \,|\, \vec{v}_L)$, in this paper we will consider only the information available in the function $P(O_n \,|\, \vec{v}_C)$. In this section we study some of the object properties that can be inferred based on contextual features. The object O_n is represented by a set of parameters $O_n = \{o_n$= category, $\vec{x}$= location, σ = scale, pose, 3D model, appearance parameters, ... $\}$. The probability function $P(O_n \,|\, \vec{v}_C)$ will introduce priors on the different object parameters. The strength of the relationship $P(O_n \,|\, \vec{v}_C)$ will depend on the nature of the contextual features used and on the object properties considered. For instance, the size of people in the image has a strong relationship with the scale of the environment and therefore one might expect that contextual features will introduce strong priors in selecting human scale. However, pose parameters will be more loosely related to context. In this paper, we consider only three object properties: object category, image location and scale: $O_n = \{o_n, \vec{x}, \sigma\}$. We apply the Bayes rule successively in order to split the PDF $P(O_n \,|\, \vec{v}_C)$ in three factors that model three kinds of context priming:

$$P(O_n \,|\, \vec{v}_C) = P(\sigma \,|\, \vec{x}, o_n, \vec{v}_C) P(\vec{x} \,|\, o_n, \vec{v}_C) P(o_n \,|\, \vec{v}_C) \tag{7}$$

The meanings of the three factors are:

- *Object priming*: $P(o_n \,|\, \vec{v}_C)$. The most likely objects given context information.
- *Focus of attention*: $P(\vec{x} \,|\, o_n, \vec{v}_C)$. The most likely locations for the presence of object o_n given context information.
- *Scale selection*: $P(\sigma \,|\, \vec{x}, o_n, \vec{v}_C)$. The most likely scales (sizes, distances) of the object o_n at different spatial locations given context information.

This decomposition of the PDF leads to factors that can be easily interpreted in terms of human behavior: the context representation activates a schema of the constituent object, then the prototypical locations and scales of the most likely objects are activated. However, any other factorization is possible as the probability graph is fully connected for the three object properties considered here. For instance, $P(o_n \,|\, \sigma, \vec{x}, \vec{v}_C) P(\sigma \,|\, \vec{x}, \vec{v}_C) P(\vec{x} \,|\, \vec{v}_C)$ will start by activating a spatial organization of the main elements (with their scales), and then at the most likely locations, different objects are primed. Of course, the different factorization make a difference only if in the implementation there is a thresholding operation that makes the computation of the subsequent terms to be performed only for the most likely object parameters. In this paper, we will consider the decomposition given in eq. (7).

Although these kinds of context priming have been shown to be important in human vision (e.g. Biederman et al, 1982), computational models of object detection typically ignore the information available from the context. From a computational point of view, context priming reduces the set of possible objects and therefore the number of features needed for discriminating between objects. It reduces the need for multiscale search and focuses computational resources into the more likely spatial locations and scales. Therefore, we propose that the first stage of an efficient computational procedure for object detection comprises the evaluation of the PDF $P(O_n \,|\, \vec{v}_C)$.

IV. Context-centered representation

One of the main problems that computational recognition approaches face in including contextual information is the lack of simple representations of context and efficient algorithms for the extraction of such information from the visual input. In fact, $\vec{v}_C$, the image information corresponding to scene context, has a very high dimensionality and it conveys information regarding all objects within the

scene. There are as many ways of breaking down the dimensionality of $\vec{v}_C$ as there are possible definitions of contextual information. For instance, one way of defining the 'context' of an object is in terms of other previously recognized objects and regions 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 (Oliva and Torralba, 2001; Torralba and Sinha, 2001a). This dispenses with the need for identifying other individual objects or regions within a scene. This is the viewpoint we shall adopt in the work presented here.

A. 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. Furthermore, the human visual system is able to analyze scenes even under degraded conditions that obscure the identities of individual objects (Schyns and Oliva, 1994; Oliva and Schyns, 1997). A few recent works have taken a different approach in which the scene is represented as a whole unity, as if it was an individual object, without splitting it into constituent parts. As we will show here, a holistic scene analysis allows representing the context in a very low dimensional space (see Oliva and Torralba, 2001 for a detailed description). Previous studies (e.g. Gorkani and Picard, 1994; Carson et al, 1997; Lipson et al, 1997; Oliva and Torralba, 2001; Szummer and Picard, 1998; Torralba and Oliva, submitted; Vailaya et al, 1998; De Bonet and Viola, 1997; Clarkson and Pentland, 2000) have shown that the elements that seem to be relevant for discrimination between different scenes are:

- The statistics of structural elements: Different structural elements (e.g., buildings, road, tables, walls, with particular orientation patterns, smoothness/roughness) compose each context (e.g., rooms, streets, shopping center). As discussed in (Oliva and Torralba, 2001), the second order statistics of natural images (encoded in the Fourier spectra) are correlated with simple scene attributes (e.g. depth) and, therefore, strongly differ between distinct environmental categories.
- The spatial organization: The structural elements have particular spatial arrangements. Each context imposes certain organization laws (e.g. for streets: road on the bottom, buildings on the sides, an aperture in the center). These different organization laws introduce spatial non-stationarities in the statistics of low-level features that provide differential signatures between scene categories.
- Color distribution: color histograms and their coarse spatial distribution provide discriminant information between scene categories. Coarse color distributions have also been shown to be relevant for scene perception by subjects (Oliva and Schyns, 2000).

Due to the relationship between objects and context categories in real-world scenes, one might expect to find a strong correlation between the objects present in the scene (their location and scale) and the statistics of local low-level features in the overall scene. As described below, we use a low dimensional holistic representation that encodes the structural scene properties. Color is not taken into account in this study, although the framework can be naturally extended to include this attribute.

B. Spatial layout of main spectral components

In order to develop a scheme for representing scene-context holistically, we have to decide on the nature of the image features to use. Experimental studies (Hubel and Wiesel, 1968) have provided evidence for the use of oriented band-pass filters (such as Gabors) in the early stages of the visual pathway. Computational studies too (Gorkani and Picard, 1994; Schiele and Crowley, 2000; Rao et al,

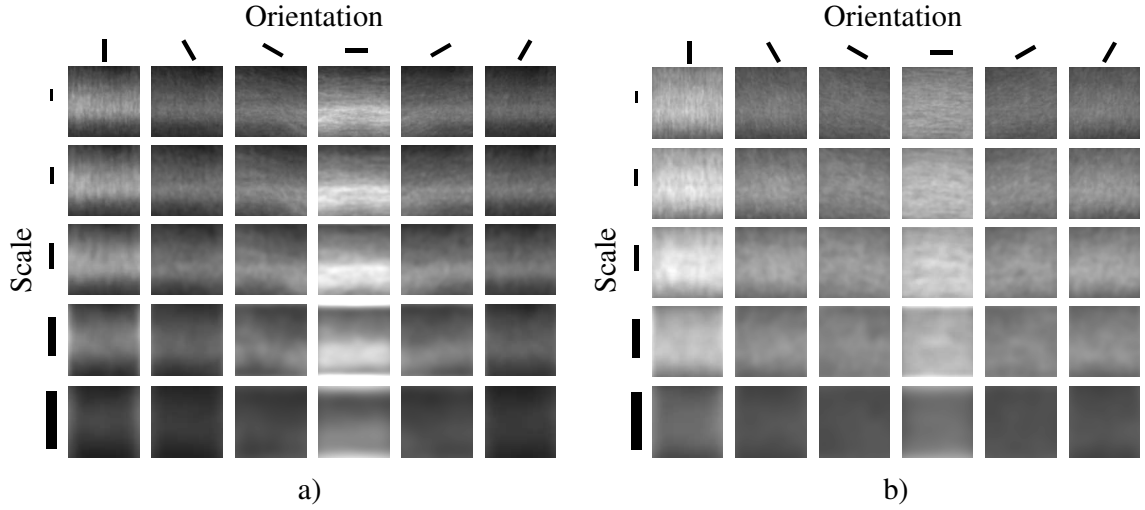

Fig. 4. Conditional average of $v(\vec{x},k)$ with respect to the presence or absence of different objects. a) $E[v(\vec{x},k)|\neg people, car]$ and b) $E[v(\vec{x},k)|people, \neg car]$. Each sub-image corresponds to the average of $v(\vec{x},k)$ for a fixed k. The index k indexes orientations and scales. Both means differ in the dominant orientations and in how the energy is distributed across coarse spatial regions.

1996; Oliva and Torralba, 2001) have found this choice of features useful for several object and scene recognition tasks. Using such features, images are encoded in a high dimensional vector: $\vec{v} = \{v(\vec{x},k)\}$ with:

$$v(\vec{x},k) = \left| \sum_{\vec{x}'} i(\vec{x}') g_k(\vec{x} - \vec{x}') \right| \tag{8}$$

where $i(\vec{x})$ is the input image and $g_k(\vec{x})$ are oriented band-pass filters defined by $g_k(\vec{x}) = g_0 e^{\|\vec{x}\|^2/\sigma_k^2} e^{2\pi j <\vec{f_k},\vec{x}>}$. In such a representation, $v(\vec{x},k)$ is the output amplitude at the location $\vec{x}$ of a complex Gabor filter tuned to the spatial frequency $\vec{f_k}$. The variable k indexes filters tuned to different spatial frequencies and orientations. An alternative representation corresponds to using $g_k(\vec{x}) = h_r(\|x\|) e^{2\pi j <\vec{f_k},\vec{x}>}$ where $h_r(\|x\|)$ is a spatial window with a constant spatial size r independent of the central frequency of the filter. This representation is equivalent to the spectrogram obtained from the amplitude of the windowed Fourier transform (WFT). In both cases, the resulting image representation encodes spatially localized structural information. For the present study we use a set of filters organized in 4 frequency bands and 6 orientations.

Studies on the statistics of natural images have shown that the statistics of low-level features, like the ones encoded by $v(\vec{x},k)$, are constrained when dealing with real-world images. The use of such regularities in the statistics finds applications in neural coding (Field, 1987), lightness and reflectance perception (Weiss, 2001; Dror, 2001), and also distortion removal (Farid, 2001) among many other applications. Statistical regularities play also a role in recognition. Low-level features statistics have different regularities when considering real-world images corresponding to different scene categories (Oliva and Torralba, 2001).

Here, we study the conditional statistics of $v(\vec{x},k)$ with respect to the presence or absence of different objects. As different objects can be found in different environments, there is a correlation between the statistics of low-level features across the scene and the objects that can be found inside. Fig. 4.a shows the conditional expectation of the features $v(\vec{x},k)$ for scenes that contain cars but no people, $E[v(\vec{x},k)|\neg people, car]$, and fig. 4.b shows $E[v(\vec{x},k)|people, \neg car]$. In both cases, only images in which the present object is smaller that 20 pixels where averaged. Each conditional expectation is obtained by averaging more than 500 images from an annotated database (see section 5 for a description of the database). As shown in fig. 4 there are large differences between both signatures. They differ in the

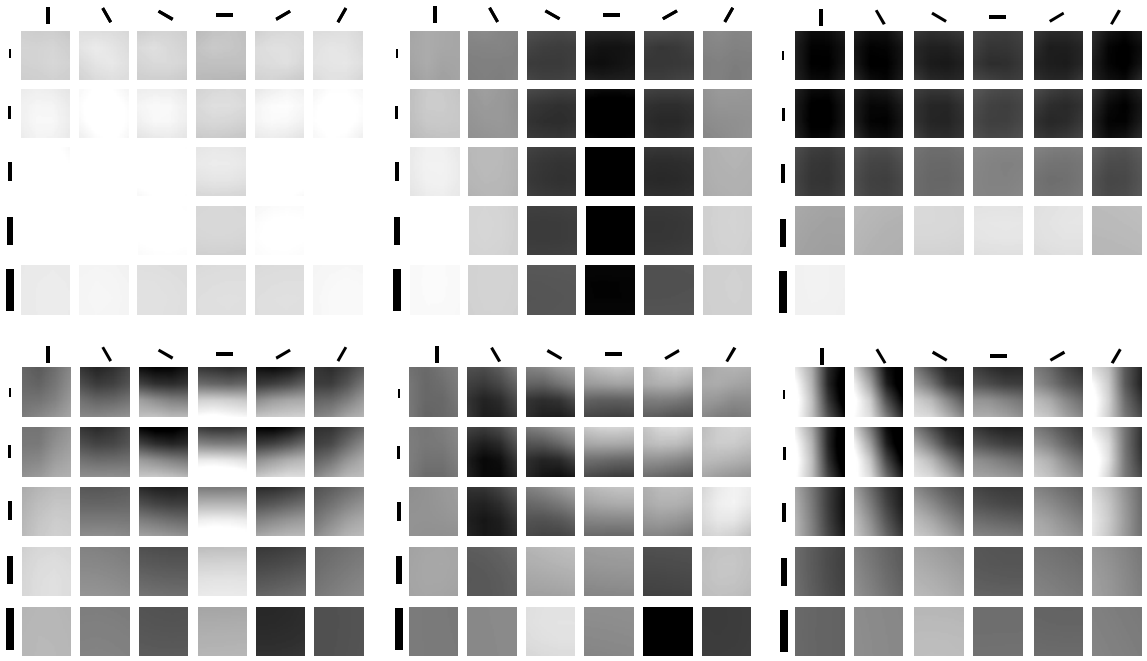

Fig. 5. Principal components $\psi_n(\vec{x},k)$ of magnitude of the output of the gabor filters. From left to right and from top to bottom, the images show the 1st, 2nd, 3td and 7th, 8th and 9th principal components. Each group of 6x5 sub-images correspond to a single principal component $\psi_n(\vec{x},k)$. The functions $\psi_n(\vec{x},k)$ incorporate both spatial and spectral information. The mean gray level correspond to the zero value. Dark pixels correspond to negative values and white pixels correspond to positive values.

dominant orientations and in how the energy is distributed across coarse spatial regions, mainly from top to bottom in the spatial domain. If these differences are stable enough across single images then they can be used for object priming as we show in the rest of the paper. The variability of the features $v(\vec{x},k)$ can be characterized by means of the principal component analysis (PCA).

We decompose the image features $v(\vec{x},k)$ into the basis functions provided by PCA:

$$v(\vec{x},k) \simeq \sum_{n=1}^{D} a_n \psi_n(\vec{x},k) \tag{9}$$

where the functions ψ_n are the eigenfunctions of the covariance operator given by $v(\vec{x},k)$. The functions $\psi_n(\vec{x},k)$ incorporate both spatial and spectral information. The decomposition coefficients are obtained by projecting the image features $v(\vec{x},k)$ into the principal components:

$$a_n = \sum_{\vec{x}} \sum_{k} v(\vec{x},k)\, \psi_n(\vec{x},k) \tag{10}$$

We propose to use the decomposition coefficients $\vec{v}_C \simeq \{a_n\}_{n=1,D}$ as context features. This approximation holds in the sense that we expect $p(O_n \,|\, \vec{v}_C) \simeq p(O_n \,|\, \{a_n\}_{n=1,D})$. D is the dimensionality of the representation. By using only a reduced set of components ($D < 64$), the coefficients $\{a_n\}_{n=1,D}$ encode the main spectral characteristics of the scene with a coarse description of their spatial arrangement. As shown in fig. 5 the first principal components encode only low resolution spatial and spectral information. The low-resolution representation, combined with the absolute value in eq. (8), provides some robustness with respect to objects arrangements that are compatible with the same scene. The representation contains information regarding the major elements that compose the scene.

In essence, $\{a_n\}_{n=1,D}$ is a holistic representation as all the regions of the image contribute to all the coefficients, and objects are not encoded individually (see Oliva and Torralba, 2001 for a detailed

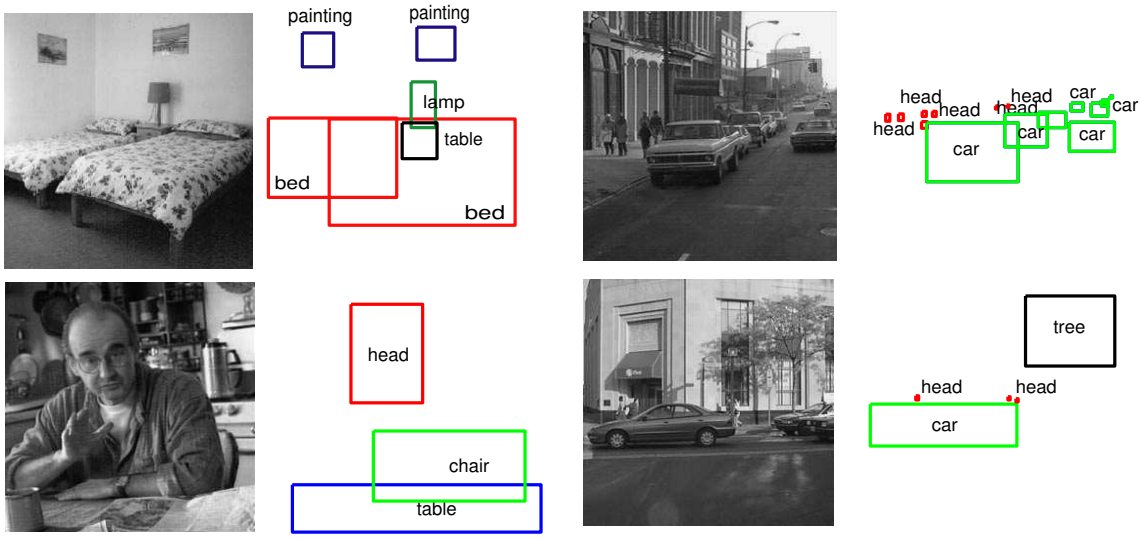

Fig. 6. Examples of images with annotations from the database used in our study.

description of the holistic representation). Note that this definition of context features differs from the one given in eq. (4) mainly because here $\vec{v}_C$ is computed from all the image measurements without discarding the ones belonging to the object O_n. This fact highlights the difference between background features and scene/context features. When the object size is small with respect to the size of the image and $\vec{v}_C$ has a low dimensionality, the scene/context features are mostly determined by the background and not by the object. However, when the target object occupies a significant portion of the image, then $\vec{v}_C \sim \vec{v}_L$.

The ability of this representation for scene recognition has been extensively explored in (Oliva and Torralba, 2001) where it is tested the performance achieved by the representation in various scene discrimination tasks and its ability to account for perceptual attributes, meaningful to observers. In (Torralba and Sinha, 2001b) the representation was shown to be robust for indoor place recognition.

V. Database annotation

In contrast with previous techniques in which the relationship between context and objects is defined by hand-written rules (Strat and Fischler, 1991), computation of the context-centered object likelihood (eq. 7) requires a learning stage in which the system learns the relationship between the contextual features and the object properties that compose the scene.

In this section we describe first the database used for training the system. The rest of the paper is devoted to the estimation of the different kinds of contextual priming formalized in eq. (7) and to showing the ability of the holistic context features to predict object properties.

The database consists of 2400 annotated 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 the purposes of the present study, four object categories where annotated: persons, vehicles, furniture and vegetation. For each picture, the annotations indicate the presence of exemplars of each object category in the scene, and the objects present are surrounded with a rectangular box. This coarse annotation allows dealing with a large number of pictures in reasonable time and provides information about the locations and approximate sizes of the objects (fig. 6).

VI. Object priming

The first factor of eq. (7), $P(o_n \,|\, \vec{v}_C)$, gives the probability of presence of the object class o_n given contextual information. If we assume that the context features vector $\vec{v}_C$ conveys enough information about the identity of the context, then there should exist strong priors on object identities, at least at the superordinate level (people, furniture, vehicles, vegetation, etc.). For instance, context-centered object priming should capture the intuition that while we do not expect to find cars in a room, we do expect to find furniture with a high probability.

The learning of the PDF $P(o_n \,|\, \vec{v}_C) = P(\vec{v}_C \,|\, o_n)P(o_n)/p(\vec{v}_C)$ with $p(\vec{v}_C) = P(\vec{v}_C \,|\, o_n)P(o_n) + P(\vec{v}_C \,|\, \neg o_n)P(\neg o_n)$ is done by approximating the in-class and out-of-class PDFs by a mixture of Gaussians:

$$P(\vec{v}_C \,|\, o_n) = \sum_{i=1}^{L} b_{i,n} G(\vec{v}_C; \vec{a}_{i,n}, \mathbf{A}_{i,n}) \tag{11}$$

where $G(\vec{v}_C; \vec{a}_{i,n}, \mathbf{A}_{i,n})$ is a multivariate Gaussian function of $\vec{v}_C$ with center $\vec{a}_{i,n}$ and covariance matrix $\mathbf{A}_{i,n}$. In this notation, the index i refers to the different clusters that model the PDF and the index n refers to the different PDFs obtained for each object class o_n. $b_{i,n}$ are the weights of each Gaussian cluster. L is the number of Gaussian clusters used for the approximation of the PDF. The model parameters ($b_{i,n}$, $\vec{a}_{i,n}$, $\mathbf{A}_{i,n}$) for the object class o_n are obtained using the EM algorithm. The learning requires the use of few Gaussian clusters ($L = 2$ yields very good performance).

The training set used for learning the PDF $P(\vec{v}_C \,|\, o_n)$ is a random subset of pictures that contain the object o_n. The training data is $\{\vec{v}_t\}_{t=1,N_t}$ where $\vec{v}_t$ are the contextual features of the picture t of the training set. The EM algorithm is an iterative procedure composed of two steps (e.g. Jordan and Jacobs, 1994; Moghaddam and Pentland, 1997; Gershnfeld, 1999):

- E-step: Computes the posterior probabilities of the clusters $h_i(t)$ given the observed data v_t. For the k iteration:

$$h_{i,n}^k(t) = \frac{b_{i,n}^k G(\vec{v}_t; \vec{a}_{i,n}^k, \mathbf{A}_{i,n}^k)}{\sum_{i=1}^{L} b_{i,n}^k G(\vec{v}_t; \vec{a}_{i,n}^k, \mathbf{A}_{i,n}^k)} \tag{12}$$

- M-step: Computes the most likely cluster parameters by maximization of the join likelihood of the training data:

$$b_{i,n}^{k+1} = \frac{\sum_{t=1}^{N_t} h_{i,n}^k(t)}{\sum_{i=1}^{L}\sum_{t=1}^{N_t} h_{i,n}^k(t)} \tag{13}$$

$$a_{i,n}^{k+1} = \frac{\sum_{t=1}^{N_t} h_{i,n}^k(t)\, \vec{v}_t}{\sum_{t=1}^{N_t} h_{i,n}^k(t)} \tag{14}$$

$$\mathbf{A}_{i,n}^{k+1} = \frac{\sum_{t=1}^{N_t} h_{i,n}^k(t)\, (\vec{v}_t - \vec{a}_{i,n}^{k+1})(\vec{v}_t - \vec{a}_{i,n}^{k+1})^T}{\sum_{t=1}^{N_t} h_{i,n}^k(t)} \tag{15}$$

The EM algorithm converges after a few iterations (< 20). The same scheme holds for the out-of-class PDF $P(\vec{v}_C \,|\, \neg o_n)$ which provides the statistical distribution of the contextual features in the set of images in which the object class o_n is not present. The probability of the object presence is approximated by $P(o_n) = P(\neg o_n) = 0.5$ (using the frequency of presence of the object-class in our database as an estimate of $P(o_n)$ does not change the results).

Figure 7 shows some typical results from the priming model on four categories of objects (o_1=people, o_2=furniture, o_3=vehicles and o_4=trees). Note that the system predicts the presence of an object based on contextual information and not on the actual presence of the object. In other words, the PDF $P(o_n \,|\, \vec{v}_C)$ evaluates the consistency of the object o_n with the context $\vec{v}_C$ and, therefore, provides information about the probable presence/absence of one object category without scanning the picture

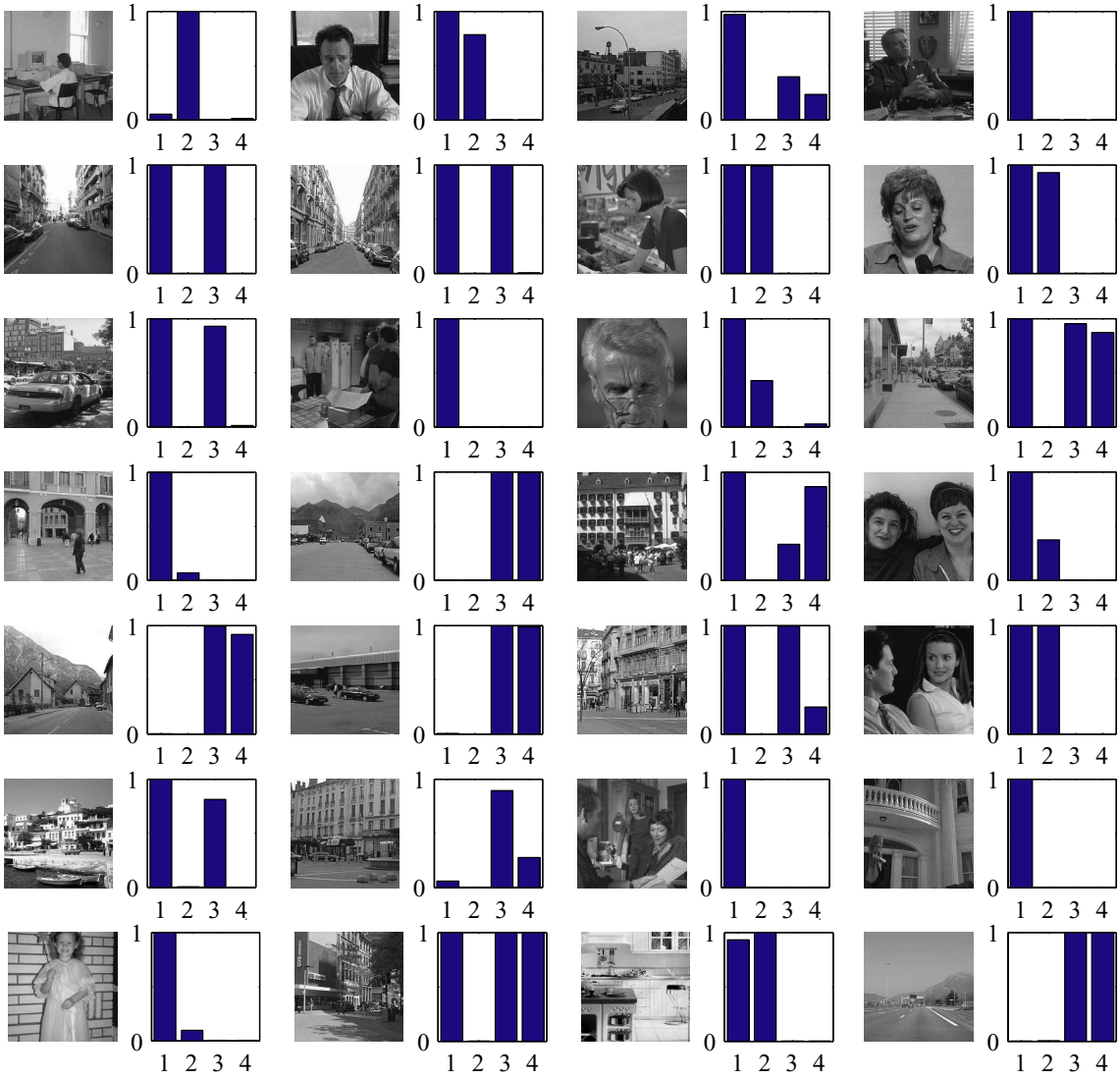

Fig. 7. Random selection of images from the test set showing the results of object priming for four superordinate object categories (o_1=people, o_2=furniture, o_3=vehicles and o_4=trees). The bars at the right-hand of each picture represent the probability $P_o(o_n \mid \vec{v}_C)$.

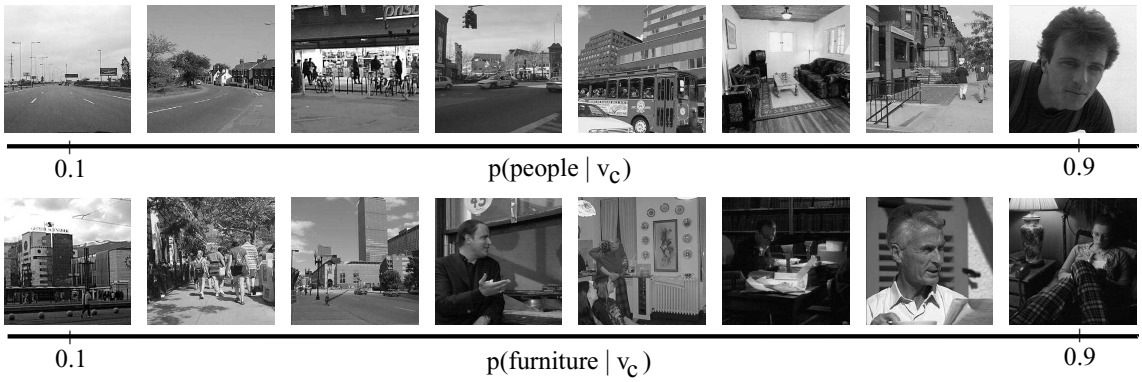

Fig. 8. Random selection of images from the test set organized with respect to the probability $P(o_n \mid \vec{v}_C)$ for o_n = people and furniture.

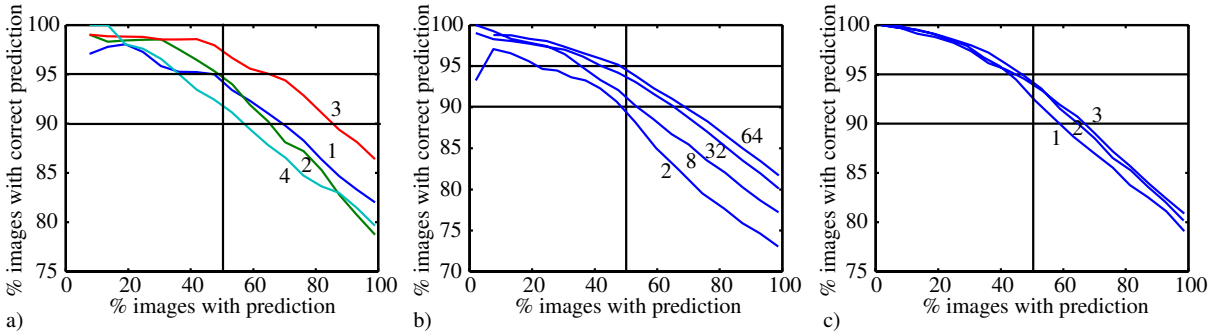

Fig. 9. Performances as a function of decision threshold th and a) target object (o_1=people, o_2=furniture, o_3=vehicles and o_4=trees), b) number of contextual features and c) number of gaussians for modeling the PDF $P(o_n \mid \vec{v}_C)$.

looking for the object. For instance, in some of the examples in fig. 7 and in fig. 11.a, the system predicts the possible presence of people based on the context (a kitchen) even if in the scene the object is missing.

Figure 8 shows a set of images organized with respect to the value of the estimated $P(o_n \mid \vec{v}_C)$ when looking for people and for furniture. The value of $P(o_n \mid \vec{v}_C)$ provides the degree of confidence that the contextual features give for deciding about the presence or absence of the object. If $P(o_n \mid \vec{v}_C) > th$ with $th \simeq 1$ then there is high confidence for deciding about the presence of the object without scanning the picture. On the other hand, if $P(o_n \mid \vec{v}_C) < 1 - th$ then the system can reliably decide that the object cannot be present in the scene represented by $\vec{v}_C$. The pictures on the extreme ends of each row are those for which the system has high confidence about the presence or absence of the target objects (people and furniture). The pictures in the middle of each row of fig. 8 represent scenes for which the system cannot reliably decide about the presence/absence of the object given the contextual features $\vec{v}_C$.

In order to study the object priming performance, we test the ability of the estimated PDF $P(o_n \mid \vec{v}_C)$ to predict the presence/absence of the object in a forced choice task as described before. By modifying the value of the threshold th we change the amount of scenes for which the system will be forced to take a decision. By setting $th = 0.5$, the system will take a decision for 100% of the images of the test set. With $th \sim 1$, the system will make a decision only for the high confidence situations. The percentage of scenes for which the system will produce high confidence predictions depends on 1) the target object and the strength of the relationship with its context, 2) the ability of the representation $\vec{v}_C$ for characterizing the scene/context and 3) the quality of the estimation of the PDF $P(o_n \mid \vec{v}_C)$. Fig. 9 summarizes the system performance as a function of these three factors. For the three graphs, the horizontal axis represents the percentage of pictures for which the system is forced to take a decision. The percentage is adjusted by changing the decision threshold th. In general, performance decreases as we force decisions for low confidence rated scenes. Fig. 9.a shows the performances for the four objects tested. The best performance was obtained for predicting the presence of vehicles (o_3) and the lowest performance corresponded to predicting the presence of vegetation (o_4). Fig. 9.b shows the results for different dimensionalities of the representation (coefficients of the PCA decomposition). Performance does not improve much when using more than 32 contextual features. Fig. 9.c shows change in performance when increasing the number of gaussians used for modeling the PDF. As shown, one gaussian already provides good results (we have also experimented with other methods such as Parzen window approximation and K-NN, and have obtained similar results). On average, when setting the threshold th in order to force decision on at least in 50% of the images, the system yields 95% correct prediction rate within this set of labeled images (with two gaussians and 32 contextual features). Prediction rate is 82% when making decisions on the entire test database.

Figure 10 shows scenes belonging to the sets defined by $P(vehicles \mid \vec{v}_C) < 0.05$ and $P(vehicles \mid \vec{v}_C) >$

Fig. 10. a) Images with low prior probability for the presence of vehicles. b) Images with high prior probability for the presence of vehicles.

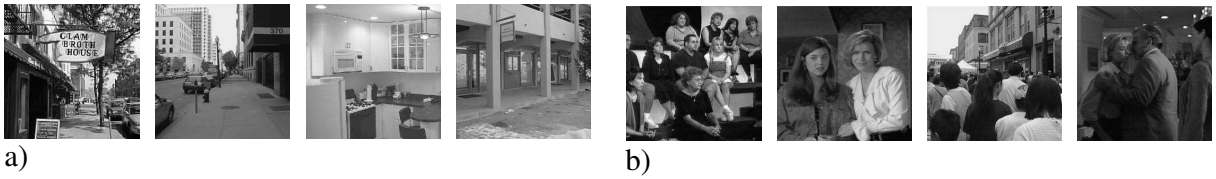

Fig. 11. Examples of scenes belonging to the sets defined by a) $P(people\,|\,\vec{v}_C) > 0.95$ and b) $P(furniture\,|\,\vec{v}_C) > 0.95$ in which the target object are missing.

0.95. 51% and 34% images of the test set belong to each group respectively. The results reveal the ability of the holistic contextual features to distinguish between different environments (Oliva and Torralba, 2001).

Figure 11 shows scenes belonging to the sets defined by $P(people\,|\,\vec{v}_C) > 0.95$ and $P(furniture\,|\,\vec{v}_C) > 0.95$ in which the target object are missing. Although these images are considered as prediction errors in terms of the results presented in fig. 9, in most of the cases, the system predictions are in agreement with the nature of the context.

In general, the introduction of other object families into the model does not require learning an increasing number of PDFs. Another way of writing the object category priming PDF $P(o_n|\,\vec{v}_C)$ is:

$$P(o_n|\,\vec{v}_C) = \sum_{i=1}^{Ncat} P(o_n|\,C_i, \vec{v}_C) P(C_i|\,\vec{v}_C) \simeq \sum_{i=1}^{Ncat} P(o_n|\,C_i) P(C_i|\,\vec{v}_C) \tag{16}$$

where $\{C_i\}_{i=1,N_{cat}}$ refers to N_{cat} non-overlapping contextual categories (for instance, street, sidewalks, room, office, etc.). The assumption is that $P(o_n|\,C_i, \vec{v}_C) \simeq P(o_n|\,C_i)$ which requires defining the correct set of contextual categories. This formulation is a dual formulation in which a context recognition step precedes object priming. It requires learning the distribution of contextual features corresponding to contextual categories instead of the presence/absence of objects categories. This formulation has the additional caveat that it requires defining appropriate contextual categories. This would be efficient only if there are less context categories than objects categories. In practical situations this may well be true. In such cases, once the probabilities $P(C_i\,|\,\vec{v}_C)$ have been learned, object priming requires the specification of the matrix $p_{n|i} = P(o_n\,|\,C_i)$ which specifies the probability of presence of the object category o_n in the context category C_i. Oliva and Torralba (2001) have shown the success of the features $\vec{v}_C$ in predicting scene categories.

Fig. 12. Examples of scenes sharing similar probabilities in the component objects. The introduction of more object categories yield more similar contexts.

The likelihood of presence of several objects categories given context information provides a signature that relates the context features to the scene category. Figure 12 shows several examples of scenes sharing the same object likelihood for the four object categories defined in our experiments. In general, scenes sharing same component objects belong to the same category.

All the inferences regarding scene and object categories can be inverted. For instance, if one object has been reliably detected, we can use the properties of the object (its location, size and category) to infer the nature of the context in which it is immersed by using $P(C_i \,|\, O_n)$. Having demonstrated the ability of holistic contextual cues to predict object categories, we next turn to studying how well such cues can facilitate object localization.

VII. Context-driven focus of attention

One of the strategies that biological visual systems use to deal with the analysis of complex real-world scenes is to selectively focus attention into the image regions that require a detailed analysis, neglecting less important regions. The goal is to focus limited computational resources into relevant scene regions. It is unclear, however, what mechanisms the visual system uses in order to make decisions about the importance of scene regions before they have been analyzed in detail.

There are several studies modeling the control of focus of attention. The most popular ones are based on low-level saliency maps (without any high-level information relative to the task or context, e.g. Itti et al, 1998; Lindeberg, 1993; Treisman and Gelade, 1980; Wolfe, 1994). Saliency maps provide a measure of the 'saliency' of each location in the image based on low-level features such as intensity contrast, orientations, color and motion. Regions that have different properties than their neighborhood are considered salient (more informative) and attract attention. The image is then explored by analyzing in detail the salient regions in the scene. These models do not take into account any high-level information (the identity of the scene) or task constraints (looking for a particular object). Other algorithms propose to include models of the target in order to account for task dependent constraints (e.g. Rao et al, 1996; Moghaddam and Pentland, 1997). But again, common to all these models is the use of features in a *local-type* or object-centered framework ignoring more high-level context information that is available in a *global-type* framework. When considering real world-scenes, it is very likely that visual search strategies and the computation of saliency maps are modulated by global high-level information related to the scene (De Graef et al 1990; Henderson and Holligworth, 1999).

In this section we propose a model of the contextual control of the focus of attention. Figure 13

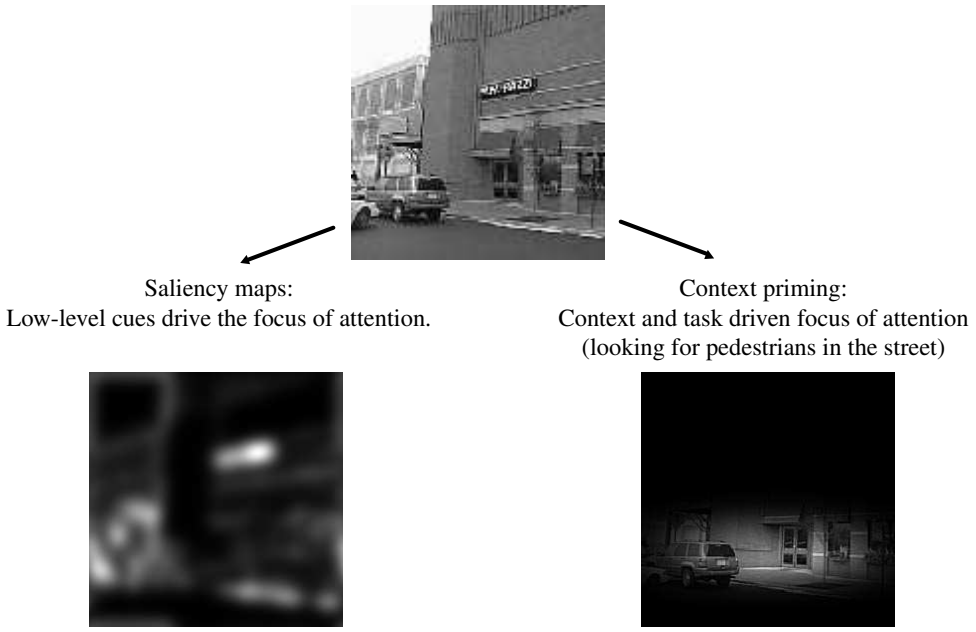

Fig. 13. Two different approaches for focus of attention. In both cases the goal is to focus computational resources into potentially relevant image regions. On the left-hand side, the focus of attention is directed toward the salient regions (in this example, saliency is simply a measure of local intensity contrast). On the right-hand side, contextual control of the focus of attention directs attention towards the sidewalk when looking for pedestrians. No local information or target models are used to drive attention here.

illustrates the differences in the control of focus of attention between a pure bottom-up saliency map (see Itti et al, 1998 for a model of low-level saliency maps) and a global context driven approach. The control of the focus of attention by contextual information is both task driven (looking for object o_n) and context driven (given global context information: $\vec{v}_C$), however, it is only based on global contextual information and does not include any model of the object (object-centered features). The context-driven approach associates (during a learning stage) the contextual features and the typical locations of the objects that compose the scene.

From an algorithmic point of view, contextual control of the focus of attention is important as it avoids expending computational resources in spatial locations with low probability of containing the target based on prior experience. It also provides criteria for rejecting possible false detections or salient features that fall outside the primed region. When the target is small (a few pixels), the problem of detection using only local features (with intrinsic target models or by saliency maps) is ill-posed. For instance, in large views of urban scenes (see figure 14), 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 as they fall outside the 'pedestrian region' (see examples of this in fig. 14).

In our framework, the problem of contextual control of the 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 ($\vec{v}_C$). It involves the evaluation of the PDF $P(\vec{x}|\, o_n, \vec{v}_C)$. For each location, the PDF gives the probability of presence of the object o_n given the context $\vec{v}_C$.

The PDF $P(\vec{x}|\, o_n, \vec{v}_C)$ is obtained via a learning stage. The learning provides the relationship between the context and the more typical locations of the objects belonging to one family. For modeling the PDF we use a mixture of gaussians (Gershnfeld, 1999):

$$P(\vec{x}, \vec{v}_C|\, o_n) = \sum_{i=1}^{M} b_{i,n}\, G(\vec{x}; \vec{x}_{i,n}, \mathbf{X}_{i,n}) G(\vec{v}_C; \vec{v}_{i,n}, \mathbf{V}_{i,n}) \tag{17}$$

Fig. 14. Focus of attention based on global context configuration. Each pair shows the original image and the image multiplied by the function $P(\vec{x}\,|\,\vec{v}_C, o_n = heads)$ to illustrate the primed regions.

The join PDF is modeled as a sum of gaussian clusters. Each cluster is decomposed into the product of two gaussians. The first gaussian models the distribution of object locations and the second gaussian models the distribution of contextual features for each cluster. The center of the gaussian distribution of object locations is written as having a linear dependency with respect to the contextual features for each cluster: $\vec{x}_{i,n} = \vec{a}_{i,n} + \mathbf{A}_{i,n}(\vec{v}_C - \vec{v}_{i,n})$.

The training set used for the learning of the PDF $P(\vec{x}, \vec{v}_C|\, o_n)$ is a random subset of the pictures that contain the object o_n. The training data is $\{\vec{v}_t\}_{t=1,N_t}$ and $\{\vec{x}_t\}_{t=1,N_t}$ where $\vec{v}_t$ are the contextual features of the picture t of the training set and $\vec{x}_t$ is the location of object o_n in the scene (we take into account only one exemplar of the multiple instances of the object in the scene). The EM algorithm is now (see Gershnfeld, 1999 for a description of the learning equations):

- E-step: Computes the posterior probabilities of the clusters $h_i(t)$ given the observed data $\vec{v}_t$ and $\vec{x}_t$. For the k-th iteration:

$$h^k_{i,n}(t) = \frac{b_{i,n}\, G(\vec{x}_t; \vec{x}^k_{i,n}, \mathbf{X}^k_{i,n}) G(\vec{v}_t; \vec{v}^k_{i,n}, \mathbf{V}^k_{i,n})}{\sum_{i=1}^{L} b_{i,n}\, G(\vec{x}_t; \vec{x}^k_{i,n}, \mathbf{X}^k_{i,n}) G(\vec{v}_t; \vec{v}^k_{i,n}, \mathbf{V}^k_{i,n})} \tag{18}$$

- M-step: Computes the most likely cluster parameters by maximization of the join likelihood of the training data:

$$b^{k+1}_{i,n} = \frac{\sum_{t=1}^{N_t} h^k_{i,n}(t)}{\sum_{i=1}^{L} \sum_{t=1}^{N_t} h^k_{i,n}(t)} \tag{19}$$

$$\vec{v}_{i,n}^{k+1} =< \vec{v} >_i = \frac{\sum_{t=1}^{N_t} h_{i,n}^k(t)\, \vec{v}_t}{\sum_{t=1}^{N_t} h_{i,n}^k(t)} \tag{20}$$

$$\mathbf{V}_{i,n}^{k+1} =< (\vec{v}_t - \vec{v}_{i,n}^{k+1})(\vec{v}_t - \vec{v}_{i,n}^{k+1})^T >_i \tag{21}$$

$$\mathbf{A}_{i,n}^{k+1} = \left(\mathbf{V}_{i,n}^{k+1}\right)^{-1} < (\vec{v} - \vec{v}_{i,n}^{k+1})\vec{x}^T >_i \tag{22}$$

$$\vec{a}_{i,n}^{k+1} =< \vec{x} - \left(\mathbf{A}_{i,n}^{k+1}\right)^T \vec{v} >_i \tag{23}$$

$$\mathbf{X}_{i,n}^{k+1} =< \left(\vec{x} - \vec{a}_{i,n}^{k+1} - \left(\mathbf{A}_{i,n}^{k+1}\right)^T \vec{v}\right)^T \left(\vec{x} - \vec{a}_{i,n}^{k+1} - \left(\mathbf{A}_{i,n}^{k+1}\right)^T \vec{v}\right) >_i \tag{24}$$

where i indexes the M clusters. The notation $<>_i$ represents the weighted average with respect to the posterior probabilities of cluster i as detailed in eq. (20). All vectors are column vectors.

Once the parameters of the joint PDF are computed, the conditional PDF P is obtained as:

$$P(\vec{x}|\, o_n, \vec{v}_C) = \frac{\sum_{i=1}^{M} b_i\, G(\vec{x}; \vec{x}_i, \mathbf{X}_i) G(\vec{v}_C; \vec{v}_i, \mathbf{V}_i)}{\sum_{i=1}^{M} b_i\, G(\vec{v}_C; \vec{v}_i, \mathbf{V}_i)} \tag{25}$$

This PDF formalizes the contextual control of the focus of attention. When looking for the object o_n, attention will be directed into the candidate regions with the highest likelihood $P(\vec{x}|\, o_n, \vec{v}_C)$ of containing the target based on the past experience of the system. The search should not be affected by locally salient features (as in fig. 13) outside the primed regions. Figure 14 shows several examples of images and the selected regions based on contextual features. In such examples the target object is a human head.

In order to better understand the behavior of the model, we can estimate the center of the region of focus of attention as:

$$(\overline{x}, \overline{y}) = \int \vec{x} P(\vec{x}|o_n, \vec{v}_C) d\vec{x} = \frac{\sum_{i=1}^{M} b_i\, \vec{x}_i\, G(\vec{v}_C; \vec{v}_i, \mathbf{V}_i)}{\sum_{i=1}^{M} b_i\, G(\vec{v}_C; \vec{v}_i, \mathbf{V}_i)} \tag{26}$$

and the width of the selected region:

$$\sigma_r^2 = \int r^2 P(\vec{x}|o_n, \vec{v}_C) d\vec{x} \tag{27}$$

with $r^2 = (x - \overline{x})^2 + (y - \overline{y})^2$ and $\vec{x} = (x, y)$.

Figures 15.a and 15.b summarize the results obtained when the target object is human heads. Figure 15.a compares the coordinate y of center of the focus of attention provided by the contextual features with respect to the average vertical location of the heads with each scene. Figure 15.b compares the x coordinate of the center of focus of attention with respect the average horizontal location of heads. It is evident from the results that global contextual features provide relevant information for the estimation of the image elevation at which faces are located. However, it does not allow the estimation of the x coordinate. This is consistent with the fact that while context places constraints on elevation (a function of ground level), it typically provides few constraints in the horizontal location of heads. This is also evident in fig. 14 where the selected regions are elongated horizontally. In general, scenes are organized along horizontal layers where the reference point is the ground level. The functionalities and the objects inside each layer (in man-made environments) are constrained by the human size. Figures 15.c and 15.d compare the location of one head in the image with respect to the average location of the rest of heads in the scene. This allows us to verify that there exists a strong correlation between the y location of heads in the scene, but the x coordinate of two heads within the same scene is decorrelated. Figures 15.c and 15.d correspond to the contextual priming provided by the objects already recognized for the detection of the remaining objects of the same category in the scene: $P(\vec{x}|o_n, O_1, O_2, ...)$.

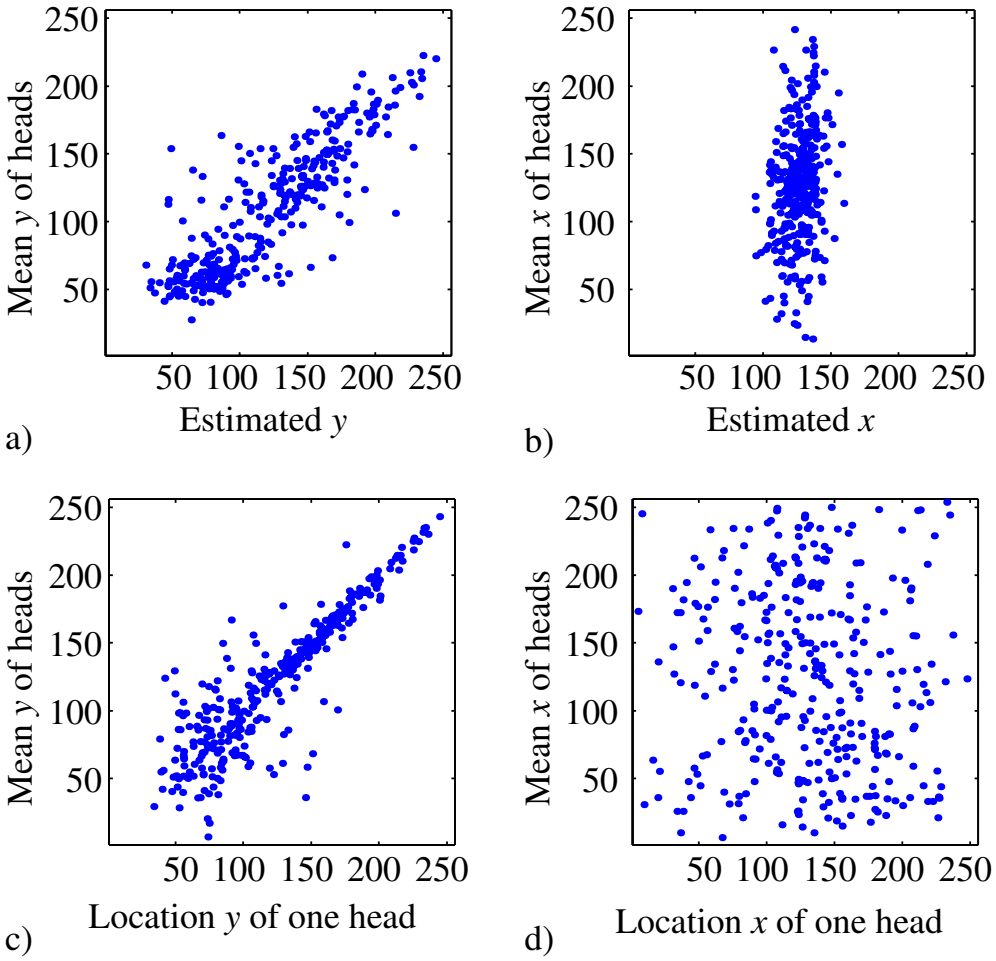

Fig. 15. Graphs (a) and (b) show respectively the comparison of the coordinates $\overline{y}$ and $\overline{x}$ of center of the focus of attention provided by the contextual features with respect to the average vertical and horizontal location of the heads in each scene. Graphs (c) and (d) compare the location of one head in the image with respect to the average location of the rest of heads in the scene.

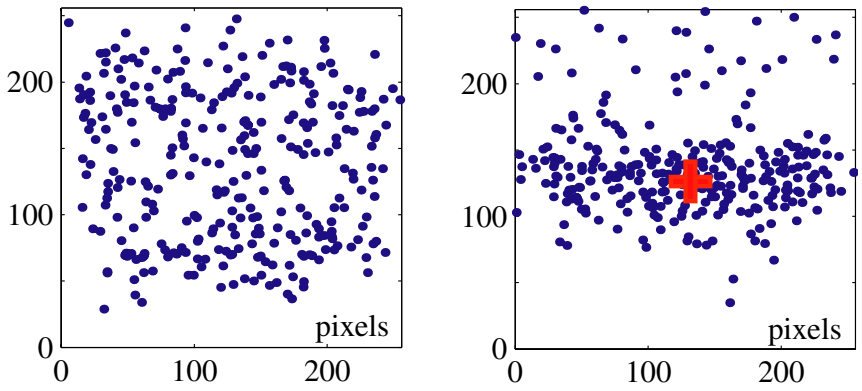

Fig. 16. a) Spatial distribution of head locations across the database. All locations are equi-probable. Therefore, an object-centered system requires exhaustive search for head detection. b) Distribution of heads with respect to the center of the focus of attention. It is evident that context provides relevant information for the estimation of the image elevation at which faces are located. However, it does not allow accurate estimation of the x coordinate. This is consistent with the fact that while context places constraints on elevation (a function of ground level), it typically provides few constraints on the horizontal location of heads.

Figure 16.a shows the distribution of the locations of the heads in the test database which is almost uniform. Therefore, in absence of additional information, all the locations in the image are equally likely to contain the target object. Figure 16.b illustrates the distribution of heads with respect to the center of the focus of attention. The success of the function $P(\vec{x}|\, o_n, \vec{v}_C)$ in narrowing the region of the focus of attention will depend on the rigidity of the relationship between the object and the context. In order to test for the reduction of the size of the search region we define the region with $P(\vec{x}|\, o_n, \vec{v}_C) > th$ with $0 < th < 1$ being a constant threshold. By adjusting the threshold th we change the size of the search region from 100% of the image size, $th = 0$, to a small image region, $th \simeq 1$. Figure 17 show the results that summarize the success of the contextual features in the reduction of the search region. For the four graphs, the horizontal axis correspond to the size of the selected image region in % and the vertical axis correspond to the percent of instances of the target object that fall inside the selected region (when there are multiple instances of the same object within a scene we only consider one randomly chosen for computing the performance). For comparison purposes, in figs. 17.a and 17.b, we also show the performances obtained when 1) contextual information is not used (in such a case, the region is selected according to the $P(\vec{x}|\, o_n)$ computed across the database) and 2) when the contextual information is provided by other instances of the object that have been already detected $(O_1, O_2, ...)$, then, the region is selected according to $P(\vec{x}|\, o_n, O_1, O_2, ...)$ that is approximated as a gaussian distribution centered on the mean location $\vec{x}_m$ of the objects $(O_1, O_2, ...)$. This provides an approximation of the upper bound on performances that can be expected from contextual information. Fig. 17.a shows performances as a function of the number of contextual features when the target object is heads, and Fig. 17.b shows performances as a function of the number of clusters used for modeling the joint PDF. The best performances are obtained with 8 clusters and 32 features. Increasing the dimensionality or the number of clusters beyond these numbers did not significantly improve the results.

The width of the selected region σ_r^2 (eq. 27) also provides a confidence measurement for the strength of the relationship between the contextual features and target location. Fig. 17.c shows the performance when selecting the 50% and 25% of the images with the lowest σ_r^2 from the test database. When considering the full test database, th needed to be set to select a region of 35% of the size of the image to guarantee that 90% of the targets will be in the selected region. When selecting the 50% of images with lowest width σ_r^2, then the target is 90% of the times within a region of size 25% of the image, and the region size becomes 20% of the image for the top 25% of the test images with the lowest σ_r^2.

Fig 17.d shows performance for the four object classes used in the study. The location of furniture and vegetation is mostly unconstrained. Cars and people (heads) show similar performance.

The aim of this section was to provide the basis for modeling the contextual control of the focus of attention based on holistic context-centered information. The procedure provides a simple framework for modeling the relationship between the scene/context and the locations of the objects that compose the scene. The model does not include any target model or local analysis and the results shows the strong constraints global scene structure provides for localizing objects. In the next section we show how context constraints also the scales of the objects that can be found inside the scene.

VIII. Context-driven scale selection

Scale selection is a fundamental problem in computational vision. Multi-scale search constitutes one of the key bottlenecks for object detection algorithms based on object-centered representations. If scale information could be estimated by a pre-processing stage, then subsequent stages of object detection and recognition would be greatly simplified by focusing the processing only onto the diagnostic/relevant scales.

Previous studies in automatic scale selection are based on bottom-up approaches (e.g. Lindeberg 1993). Similar to approaches of spatial focus of attention, scale selection has been based on measurements of the saliency of low-level operators across spatial scales. For instance, Lindeberg (1993)

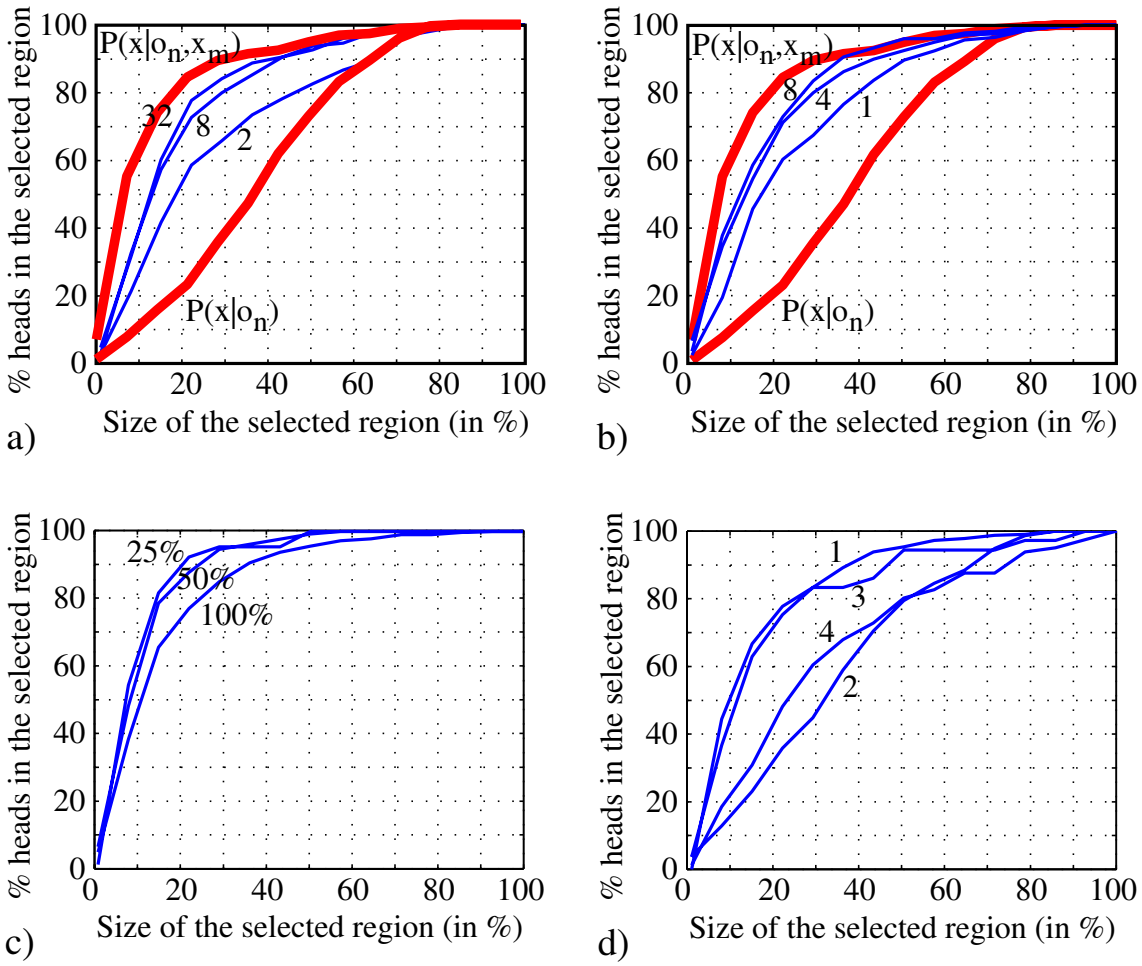

Fig. 17. Quantifying the ability of contextual features to reduce the size of the search region as a function of a) number of contextual features ($D = 2, 8, 32$). Graph (c) shows performances for the top 25% and 50% of images with the lowest σ_r^2. Graph (d) compares performances across the four object classes (1-people, 2-furniture, 3-vehicles and 4-vegetation).

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. However, when looking for particular objects (like pedestrians in a street scene), the target object will not always appear as a salient blob or even a well-defined shape.

Here we show that the holistic context features provide a strong cue for scale selection for the detection of high-level structures as objects. In the model we propose, automatic scale selection is performed by the PDF $P(\sigma \,|\, o_n, \vec{v}_C)$. For simplicity, we have assumed that the scale is independent of position: $P(\sigma \,|\, \vec{x}, o_n, \vec{v}_C) \simeq P(\sigma \,|\, o_n, \vec{v}_C)$. This PDF relates the typical scales σ (image size in pixels) of the object o_n with the contextual features $\vec{v}_C$. As the scene structure restricts the possible positions and distances at which objects can be located we can expect that the PDF $P(\sigma \,|\, o_n, \vec{v}_C)$ provides relevant information for scale selection. This is illustrated in fig. 18 which shows the conditional average of the output of Gabor filters at different scales and orientation when the scenes contain people at three different scales. In this section we show that the differences between the signatures of each scale are stable enough to provide reliable object scale priming.

The model for the conditional PDF $P(\sigma \,|\, \vec{x}, o_n, \vec{v}_C)$ is similar to the one used for modeling the focus

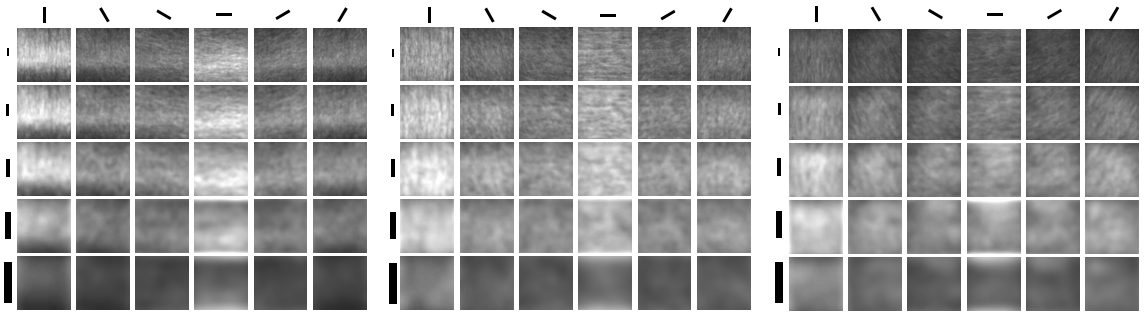

Fig. 18. Conditional average of $v(\vec{x}, k)$ for images that contain people at three different scales 5, 25 and 125 pixels (the images are 256^2 pixels is size) . No other constrains are impose for the other objects.

of attention (eq. 25):

$$P(\sigma | o_n, \vec{v}_C) = \frac{\sum_{i=1}^{M} b_i \, G(\sigma; \sigma_i, \mathbf{S}_i) G(\vec{v}_C; \vec{v}_i, \mathbf{V}_i)}{\sum_{i=1}^{M} b_i \, G(\vec{v}_C; \vec{v}_i, \mathbf{V}_i)} \tag{28}$$

with $\sigma_i = a_i + \vec{A}_i^T(\vec{v}_C - \vec{v}_i)$. The model parameters are $(b_i, a_i, \vec{A}_i, \mathbf{S}_i, \vec{v}_i, \mathbf{V}_i)_{i=1,M}$ which are obtained after a learning stage. The learning stage is performed by means of the EM algorithm as detailed in the precedent section (eqs. 19 to 24) using the database of annotated images. For the examples provided in this section we have focused in the prediction of the size of human heads ($o_n = heads$) in the scene. We estimated the scale σ as being the mean height H of the heads present in the picture (in logarithmic units): $\sigma = \log(H)$, with H given in pixels. Head height, which refers to the vertical dimension of a square box surrounding a head in the image, is mostly independent of head pose (variations in pose are mostly due to horizontal rotations). In the case of human heads, changes in pose due to horizontal rotations are unconstrained by contextual information (Torralba and Sinha, 2001c).

The preferred scale ($\overline{\sigma}$) given context information ($\vec{v}_C$) is estimated as the conditional expectation:

$$\overline{\sigma} = \int \sigma P(\sigma | o_n, \vec{v}_C) d\sigma = \frac{\sum_{i=1}^{M} \sigma_i \, b_i G(\vec{v}_C; \vec{v}_i, \mathbf{V}_i)}{\sum_{i=1}^{M} b_i \, G(\vec{v}_C; \vec{v}_i, \mathbf{V}_i)} \tag{29}$$

and the variance of the estimation is σ_h^2:

$$\sigma_h^2 = \int (\sigma - \overline{\sigma})^2 P(\sigma | o_n, \vec{v}_C) d\sigma \tag{30}$$

The model reaches maximal performance with as few as $M = 4$ clusters. Fig. 19 shows a random selection of images from the entire test set with the expected head size estimated with eq. (29). The square box indicates the estimated height and the segment at the right-hand side indicates the real mean height of the heads in the picture. The results are summarized in fig. 20.a. The estimated scale, $\overline{\sigma}$, is compared with respect to the mean scale of heads in each scene, σ_m. For 81% of the images, the real scale was inside the range of scales $\sigma_m \in [\overline{\sigma}/\alpha, \overline{\sigma} \cdot \alpha]$ with $\alpha = 2$. For comparison, we also show the scale priming for the detection of heads provided that one head has already been detected reliably. This provides an upper bound of the constraints existing between the scale of one object (heads) and its context. For 90% of the images the scale of one head, σ_1, selected at random among the multiple instances in each scene, was within the range of scales $\sigma_1 \in [\sigma_m/\alpha, \sigma_m \cdot \alpha]$, with σ_m is given by the mean scale of the rest of heads in the scene and $\alpha = 2$ (Fig. 21.a).

Fig. 21.a summarizes the results for different scale ranges α when varying the number of contextual features ($M = 4$). The results show that, in order to guarante that 90% of the heads are within the

Fig. 19. Results for scale selection given global context information for random selection of results from the test set. The size of the square box corresponds to the expected height of heads given the holistic contextual features. The line at the right hand indicates the real height of the heads when they are present in the image.

scales explored, we have to explore the range of scales given by $[\overline{\sigma}/\alpha, \overline{\sigma}\cdot\alpha]$ with $\alpha = 2.4$ and $\overline{\sigma}$ given by eq. (29). It has to be noted that, when no contextual information is taken into account, it is necessary to explore the range of scales $[\overline{\sigma}/\alpha, \overline{\sigma}\cdot\alpha]$ with $\alpha = 7.3$ in order to guarantee that 90% of the heads are within the scales explored given the variability within our database.

The variance of the estimation given by eq. (30) provides a confidence measurement for the scale priming and can be used for reducing the range of possible scales to explore for high confidence contexts. Fig. 21.b shows the results when selecting the 50% and the 25% of the images with the lowest σ_h^2 among the scenes used for the test. For the selected 50% images, it is necessary to explore the range of scales $[\overline{\sigma}/\alpha, \overline{\sigma}\cdot\alpha]$ with $\alpha = 1.7$ in order to guarantee that 90% of the heads are within the scales explored.

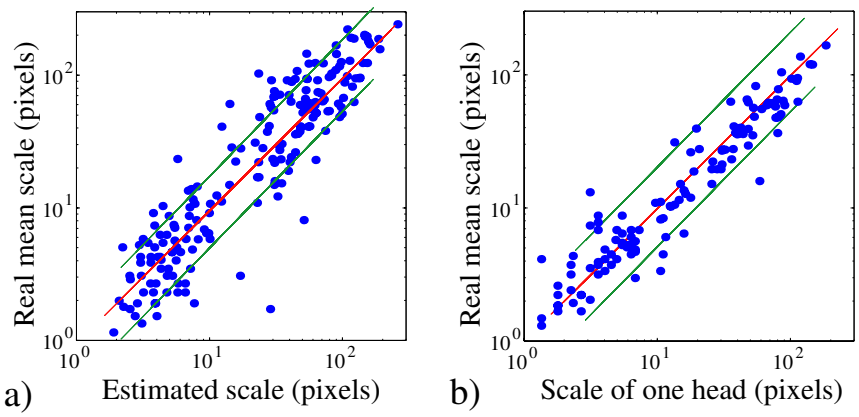

Fig. 20. a) Results for scale selection given global context information. b) Scale selection given the size of one of the objects in the picture.

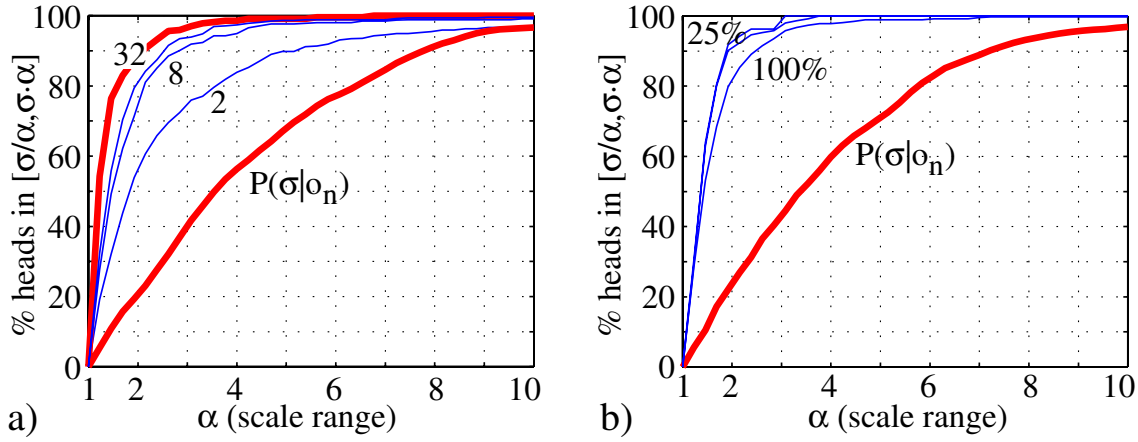

Fig. 21. The graphs show the percent of heads the are within the range of scales given by the interval $[\overline{\sigma}/\alpha, \overline{\sigma} \cdot \alpha]$ with varying α. σ is the selected scale for each image. a) Performance of scale selection as a function of the number of contextual features (2, 8 and 32). b) Scale selection performance when considering the 100%, 50% or 25% of the images of the test set according to the confidence measure σ_h^2.

The relative size (σ) of an object inside an image, depends on both the relative image size of the object at one fixed distance and the actual distance D between the observer and the object. Fig. 22 shows a set of images sorted according to the estimated scale of heads inside the image. The organization is correlated with the size of the space that the scene subtends (Torralba and Oliva, submitted).

IX. Conclusion

There are strong constrains in the statistical distribution of objects and environment in real-world scenes. Furthermore, real-world scene pictures have strong regularities of simple pixel statistics like the ones captured by linear filter outputs. Both statistical regularities, the distribution of objects and the statistics of low-level features, are linked. In particular, we showed that there are differential regularities when conditioning the statistics with respect to the presence/absence of objects and their properties. The study of such conditional regularities provides the basis for the contextual priming framework developed in this paper.

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 that captures low-resolution spatial and

Fig. 22. Each row represents a set of 8 pictures sorted according to the predicted size of human heads in the scene. In such an approach, depth is provided by familiar context.

spectral information of the image. The use of the statistical framework provides also a simple way for giving confidence measurements for the contextual priming. The strength of the contextual priming $p(O_n \,|\, \vec{v}_C)$ varies from one image to another.

However, several interesting issues remain open. These include 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] Biederman, I. 1987. Recognition-by-components: A theory of human image interpretation. *Psychological Review*, 94, 115-148

[3] 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.

[4] Clarkson, B., and Pentland, A. 2000. Framing through peripheral vision. *Proc. IEEE International Conference on Image Processing*, September 10-13, 2000 in Vancouver, BC.

[5] Chun, M. M., and Jiang, Y. 1998. Contextual cueing: Implicit learning and memory of visual context guides spatial attention. *Cognitive Psychology*, 36, 28-71.

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

[7] De Graef, P., Christiaens, D., and d'Ydewalle, G. 1990. Perceptual effects of scene context on object identification. *Psychological research*, 52:317–329.

[8] Dror, R., Adelson, T., and Willsky, A. 2001. Surface reflectance Estimation and natural illumination statistics. *Proc. of IEEE workshop on Statistical and Computational Theories of Vision*, Vancouver, CA, July 2001.

[9] Farid, H. 2001. Blind inverse gamma correction. *IEEE transactions on image processing*, in press

[10] Field, D. J. 1987. Relations between the statistics of natural images and the response properties of cortical cells. *Journal of Optical Society of America*, 4, 2379-2394

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

[12] 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.

[13] Haralick, R. M. 1983. Decision making in context. *IEEE Trans. Pattern Analysis and Machine Intelligence*, Vol. 5: 417–428.

[14] Henderson, J.M., and Hollingworth, A. 1999. High level scene perception. *Annual Review of Psychology*, 50, 243-271.

[15] Hubel, D. H., and Wiesel, T. N. 1968. Receptive fields and functional architecture of monkey striate cortex. *Journal of Physiology*, 195, 215-243.

[16] Intille, S. S., and Bobick, A. F. 1995. Closed-world tracking. *Fifth International Conference on Computer Vision*, pp. 672–678. Los Alamitos.

[17] 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.

[18] Jordan, M. I., and Jacobs, R. A. 1994. Hierarchical mixtures of experts and the EM algorithm. *Neural Computation* 6:181–214.

[19] Koch, C., and Ullman, S. 1985. Shifts in visual attention: towards the underlying circuitry. *Human Neurobiology*, 4, 219-227.

[20] Jepson, A. Richards, W., and Knill, D. 1996. Modal structures and reliable inference. *Perception as Bayesian Inference*, eds. D. Knill and W. Richards, Cambridge Univ. Press, pp. 63-92.

[21] 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.

[22] 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.

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

[24] Oliva, A., and Schyns, P. G. 1997. Coarse blobs or fine edges? Evidence that information diagnosticity changes the perception of complex visual stimuli. *Cognitive Psychology*. 34:72-107

[25] Oliva, A., and Schyns, P. G. 2000. Diagnostic color blobs mediate scene recognition. *Cognitive Psychology* , 41:176–210.

[26] 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.

[27] Palmer, S. E. 1975. The effects of contextual scenes on the identification of objects. *Memory and Cognition* , 3:519–526.

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

[29] Potter, M. C. 1975. Meaning in visual search. *Science*, 187:965-966.

[30] 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.

[31] Rensink, R. A., O'Regan, J. K., and Clark, J. J. 1997. To see or not to see: the need for attention to perceive changes in scenes. *Psychological Science*, 8:368-373

[32] Ripley, B. D. 1996. *Pattern recognition and neural networks.* Cambridge University Press.

[33] Schiele, B., and Crowley, J.L. 2000. Recognition without Correspondence using Multidimensional Receptive Field Histograms. *Int. Journal of Computer Vision*, 36(1):31–50.

[34] Schyns, P.G., & Oliva, A. 1994. From blobs to boundary edges: evidence for time and spatial scale dependent scene recognition. *Psychological Science*, 5:195-200.

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

[36] Strat, T. M., and Fischler, M. A. 1991. Context-based vision: recognizing objects using information from both 2-D and 3-D imagery. *IEEE trans. on Pattern Analysis and Machine Intelligence*, 13(10): 1050-1065.

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

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

[39] Torralba, A., and Sinha, P. 2001a. Statistical context priming for object detection. *IEEE Proc. Of Int. Conf in Comp. Vision*, Vol 1:763–770.

[40] Torralba, A., and Sinha, P. 2001b. Recognizing indoor scenes. *AI Memo* 2001-015, July 2001.

[41] Torralba, A., and Sinha, P. 2001c. Contextual modulation of target saliency. NIPS 2001.

[42] Torralba, A., and Oliva, A. Depth perception from familiar structure. Submitted.

[43] Treisman, A., and Gelade, G. 1980. A feature integration theory of attention. *Cognitive Psychology*, Vol. 12:97–136.

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

[45] Weiss, Y. 2001. Deriving intrinsic images from image sequences. *IEEE Proc. Of Int. Conf in Comp. Vision*, Vol 2:68–75.

[46] Wolfe, J. M. 1994. Guided search 2.0. A revised model of visual search. *Psychonomic Bulletin and Review*, 1:202-228