Multimed Tools Appl (2011) 51:303–339
DOI 10.1007/s11042-010-0631-y

Context modeling in computer vision: techniques, implications, and applications

Oge Marques · Elan Barenholtz · Vincent Charvillat

Published online: 10 November 2010

Abstract In recent years there has been a surge of interest in context modeling for numerous applications in computer vision. The basic motivation behind these diverse efforts is generally the same—attempting to enhance current image analysis technologies by incorporating information from outside the target object, including scene analysis as well as metadata. However, many different approaches and applications have been proposed, leading to a somewhat inchoate literature that can be difficult to navigate. The current paper provides a 'roadmap' of this new research, including a discussion of the basic motivation behind context-modeling, an overview of the most representative techniques, and a discussion of specific applications in which contextual modeling has been incorporated. This review is intended to introduce researchers in computer vision and image analysis to this increasingly important field as well as provide a reference for those who may wish to incorporate context modeling in their own work.

Keywords Computer vision · Object recognition · Objects in context · Context modeling

1 Introduction

This paper surveys recent work in the area of context modeling in computer vision. It presents a structured overview of the most representative context modeling

O. Marques (✉) · E. Barenholtz
Florida Atlantic University, 777 Glades Road, Boca Raton, FL 33431, USA
e-mail: omarques@fau.edu

E. Barenholtz
e-mail: elan.barenholtz@fau.edu

V. Charvillat
ENSEEIHT, 2 Rue Charles Camichel, 31000 Toulouse, France
e-mail: vincent.charvillat@enseeiht.fr

techniques available in the literature and discusses their applications—especially in object detection, localization, and recognition—and implications for future research in related topics.

The primary goal of this paper is to provide a user-friendly, readable, and broad overview of the field of context modeling to readers who are not yet familiar with this relatively new topic in computer vision (and its many recent developments). Hence, the paper compiles the most representative efforts in the field and organizes the knowledge of the topic in a way that should provide perspective and allow the reader to make informed choices as to where to learn more about the topic and find associated research resources.

The chief motivation for preparing this paper was the growing interest in the topic of "objects in context" in both human and computer vision during the past ten years. On the human-vision research side, the role of context in object recognition continues to be a topic of intense study, about which many subtle details (and their underlying mechanisms) remain to be discovered and well understood. From a computer vision perspective, there seems to be general agreement among researchers and practitioners that the time is ripe for solutions that somehow model and leverage the role of context on classical computer vision tasks, particularly object recognition.

The paper is structured as follows: Section 2 discusses the importance of context in human vision and provides an overview of the most representative research efforts and findings from neuroscience, cognitive psychology, and associated fields; Section 3 presents background information on early efforts to model context in computer vision solutions as well as several ways to classify different types of context available in the literature; Section 4 discusses in detail the prominent role of spatial knowledge in contextual reasoning; Section 5 reflects upon the potential implications of such research efforts to the future of computer vision; and Section 6 presents concluding remarks. The paper also includes an Appendix—targeted at readers who are interested in doing research in this field—which contains practical information about research groups, datasets and open-source code related to context modeling in computer vision.

2 The importance of context in human vision

This section presents an overview of the role of context in human visual processes, particularly object recognition. It summarizes key studies and insights on the role of context in human vision and presents a list of open questions that are driving the research efforts in this field.

2.1 Background

The human ability of detecting and recognizing objects and performing a broad range of visual tasks in a wide variety of situations, despite considerable amount of clutter, occlusions, changes in illumination and viewpoint, is a remarkable trait of our visual system, one that is still far for being completely understood, modeled, or matched by computer vision solutions. The process of *object recognition*—the ability to assign a label, name, or category to an object in a scene, based on the optical information available from such scene—has been the subject of intense study in human vision research.

Object recognition is inherently complex and has been investigated from the viewpoints of neurophysiology, behavioral psychology, brain imaging studies, etc. While the past several decades have seen significant progress concerning the nature of the *input* to the human visual system, which seems to consist of a diverse set of primitive visual features (e.g., colors, intensities, edges, orientations), the means by which this input is integrated and matched to a stored representation remains a matter of controversy (see [86] for a recent review). That is, the puzzle of human object recognition remains largely unsolved.

The majority of research on human object recognition has considered objects shown in isolation of any surrounding scene. However, more recently there has been an influx of compelling evidence that object recognition doesn't happen in isolation, i.e., the process of recognizing one object in a scene can be influenced by the presence of other objects as well as by the overall *context* of the scene. Contrary to visual search experiments, in which the target is surrounded by *distractors*—a case in which the context hinders performance of the task—, in most real-world object recognition tasks, the context provides a rich source of information that can help improve the performance of the task.

Figure 1 shows an example of object recognition task in which the context surrounding the object of interest plays a significant role in the recognition process. The image on the left is virtually impossible to recognize in isolation. However, the same image, shown in context on the right (circled) is easy to identify for our visual system. The kind of degradation shown in Fig. 1—and the potential role of context in overcoming it—is not relegated to artificially manipulated images but is highly pervasive, occurring under numerous 'real world' conditions such as poor illumination, distant viewing, peripheral viewing and occlusion by other objects. In many of these cases, the visual system appears to use context to overcome poor quality of the target image.

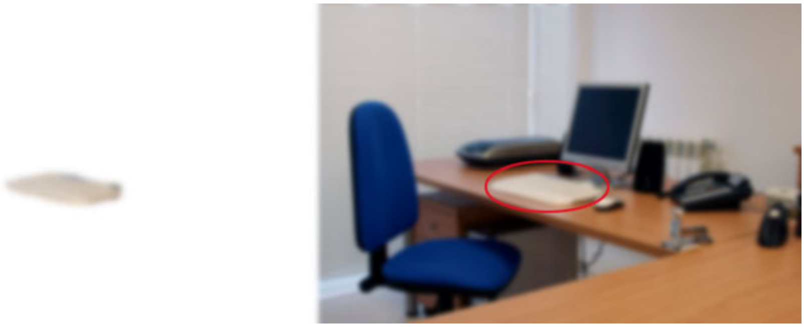

Fig. 1 An object viewed in isolation is unrecognizable (*left*) while it can be readily identified in its contextual scene (*right*)

We conclude this subsection with an operational definition of context: in this paper, we adopt and expand the definition provided by [122] and call *context* "any information that might be relevant to object detection, categorization and classification tasks, but not directly due to the physical appearance of the object, as perceived by the image acquisition system." Let us examine this definition in more detail, since it has guided—to some degree—the way this paper was written and structured. Starting from the phrase "any information", it usually means visual information available within the scene but outside the boundaries of the object of interest; in some cases, it also includes non-visual information, such as geographical data (e.g., location at which the picture was taken, expressed in GPS coordinates) , annotation data (e.g., tags added to the image by the user), and others (e.g., camera settings, weather information, etc.). Regarding visual tasks, most of the paper is devoted to object detection, categorization and classification tasks, but in Section 5.2 we expand the field to include other applications of context modeling as well. Finally, the "physical appearance of the object, as perceived by the image acquisition system" refers to the object's visual contents and features, as encoded by the human visual system (HVS) or a set of image processing and computer vision algorithms, which indirectly alludes to the fact that we are fully aware of the fundamental problem of vision—the ability to perceive a rich 3D world from a series of incomplete 2D projections, each of which could have been created by infinitely many variations of 3D scenes. However, the goal of this paper is to consider how context may facilitate processing based on currently available image processing techniques, even without first solving some of the basic problems in the field.

2.2 Contextual priming in human object detection and recognition

As mentioned above, within the human vision literature, the majority of theoretical research on object detection and recognition has followed Marr's [74] program in which identification proceeds in a bottom-up fashion, based on the locally visible properties of individual objects. However, a substantial degree of empirical work has considered the role of context in facilitating visual object recognition. For example, a large number of studies have reported an effect in which objects presented within an appropriate contextual setting are recognized more rapidly than those viewed in an inappropriate context. Typically, these studies use a paradigm in which an image or drawing of a complex scene (such as a parking lot) is displayed quickly after which subjects must decide whether a particular target object was present in the image or not. Overall, these studies suggest a facilitatory effect of context on object identification that can occur at two basic levels: *semantic* (e.g., a tractor and a barn can both appear in a farmer scene, but an octopus is not consistent with that type of scene [11–13, 26, 44, 85, 94], and *spatial relations* (e.g., a patch of sky is expected to appear above a patch of grass) [11, 28, 54, 94]. For example, in a well-known series of experiments, Biederman [11] examined performance in a detection task as a function of the number of 'violations' between a target object and the scene in which it was (briefly) presented. These violations included unexpected spatial relations, based on the size or position of the target object relative to other objects in the scene as well as violations in the semantic relations, based on the expected likelihood of a particular object appearing a specific scene. Overall, he found that as violations across these categories increased, performance in the detection task declined. However, the extent and nature of contextual facilitation of object recognition in this and other studies remain controversial with the majority arguing that the advantage reflects an advantage in perceptual or cognitive processing of the target object while others believe it reflects a 'response bias' in which subjects are guessing the correct answer *after* perceptual processing has been completed [52, 61]. It is important to note that all of these studies use stimuli that are fully recognizable even without context, making the nature of the role of context in performing recognition difficult to identify. Indeed, a number of the above studies employed a paradigm in which context does not, by design, provide any task-relevant information—for example cases in which subjects must decide which of two objects (both of which are consistent with a contextual scene) were present in the scene [5, 61]; under these circumstances, the role of context, if any, is highly indirect.

Contextual information has also been found to improve performance in *detecting* a target object amidst high degrees of visual clutter–that is, during visual search. A number of studies have found that people learn stable spatial relationships between objects and their respective contexts leading to a reduction in search for the target object. This phenomenon, referred to as 'contextual cueing', has been demonstrated both for standard letter-grid visual search stimuli, such as searching for a 'T' in a field of 'L's [22–24, 43] as well as realistically rendered 3-dimensional environments [14, 15]. In general, contextual cueing has been thought to depend on more efficient allocation of attention to probable regions of the scene. However, this interpretation has recently been challenged by studies which suggest that contextual cueing may depend on perceptual/decision processes, similar to the kinds of contextual facilitation effects described above, rather than attentional guidance [70]. Other recent research has suggested that both types of phenomena may be present. Note that standard contextual cueing experiments depend on specific target/context relations (e.g., a letter's location in a grid) learned during the course of an experiment. However, people can also learn more general scene/object relations based on their experience in the real world. Using eye-tracking during visual search of naturalistic scenes, Hidalgo-Sotelo et al. [53] found that the allocation of early fixations is guided by prior knowledge of the locations of objects within a general scene category (e.g., fixating at street level to find people) while the use of specific scene/object relations learned across repetitions (i.e. standard contextual cueing) appears afterwards, perhaps reflecting a different mechanism.

From a cognitive neuroscience perspective, a number of recent studies have suggested that the brain may encode contextual information, based either on associations between objects and scenes as well as the relations between semantically related objects. Bar and Aminoff [7] and Bar [6] found that viewing individual objects that are generally strongly associated with a particular context (e.g., a stove, which is associated with a kitchen) elicited strong responses in brain areas (particularly the parahippocampal cortex, or PHC) that are believed to be involved in encoding locations and spatial landmarks [6]. More recently, Gronau et al. [46] found that PHC activation was specifically elicited when objects were shown in their appropriate spatial relation to other objects (e.g., a mirror above a dresser). These findings have led some researchers to suggest that the PHC is specifically involved in encoding contextual relations. However, these findings are preliminary and the precise role of context in individual object recognition has still not been determined.

2.3 Contextual facilitation in recognizing degraded images

There is, in summary, a large body of evidence suggesting that contextual information impacts the efficiency of object detection and recognition tasks and “a general consensus that objects appearing in a consistent or familiar background are detected more accurately and processed more quickly than objects appearing in an inconsistent scene [81]”. However, as noted, virtually all of the previous research concerning the role of context on object recognition has considered visual stimuli in which the target object is fully recognizable to human observers in isolation (i.e. even without context). The role of context under these conditions is indirect and thus has often been difficult to characterize. An important type of contextual facilitation in human vision not addressed by these studies is in the recognition of *degraded* stimuli. Under these circumstances, context can provide direct information not available from the image of the target object itself. For example, Selfridge [99] produced a well-known demonstration in which letters made ambiguous by artificial ‘ink blotches’ covering critical features, may be identified based on the context of the word in which they appear. A similar finding was described by Bar and Ullmann [8], in which segments of stylized drawings could not be identified in isolation of the rest of the image. Perhaps the most convincing source of evidence, however, comes not from experiments using artificial stimuli but simple observations of human performance for realistic images. One of the most popular and compelling examples of the role of context in object recognition under poor conditions has been provided by Antonio Torralba [110] and became known as “the multiple personalities of a blob” (Fig. 2). In these images, the same gray blob can be interpreted as a plate, bottle, cell phone, car, pedestrian, or shoe, depending on the context. (Each

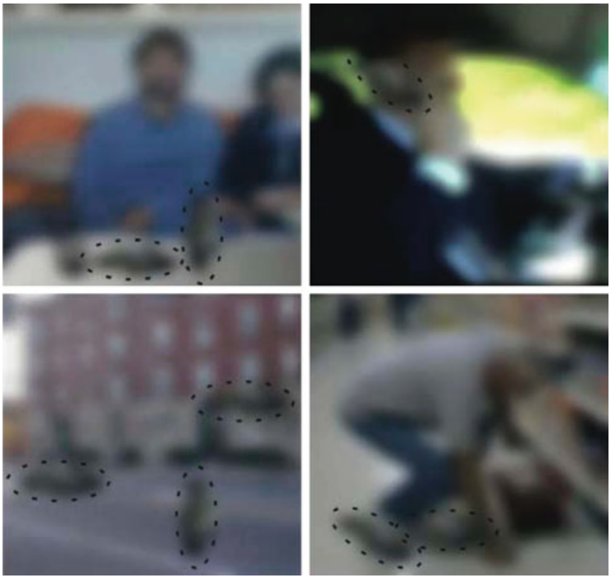

Fig. 2 The multiple personalities of a blob (from [113])

circled blob has identical pixels, but in some cases has been rotated.) Interestingly, recent research has demonstrated that this type of contextual facilitation can cause degraded stimuli to evoke a similar brain response in object-specific brain areas as non-degraded image, given the appropriate contextual surrounding [25]. This may suggest that context-facilitated object recognition and 'standard' recognition depend on the same neural mechanisms and may not represent distinct cognitive processes.

From an experimental standpoint, very little work has addressed the role of context as a source of disambiguation of degraded stimuli. Recently, Barenholtz [9] compared performance in a recognition comparing cases in which a target object was shown in isolation, where a target object was shown in a recognizable setting (e.g., a kitchen) unfamiliar to the subject and cases in which the target was shown in the subject's own home. This study used a 'pixelation' method in which the selected image region was divided into a grid of equally sized, square checks each of which contained the average value (color and luminance) of all of the screen pixels contained within its boundaries (Fig. 3). The size of the check determined the resolution of the image and was used as a measure of 'information' in the image. Using this method, Barenholtz found that the effect of context was profound: in the generic context condition (i.e. when the context was not familiar to the subject), people required about four times less visual information while in the condition in which the context was the subject's own home, they required almost 20 times less information! These results suggest that context can itself provide a great deal of information, over and above the appearance of the target object, that may be used for recognition.

While both the behavioral and neurophysiological studies described above used artificial forms of degradation, the human visual system must contend with poor image quality under numerous 'real world' conditions such as poor illumination, distant viewing, peripheral viewing and occlusion by other objects. In addition, visual identification often takes place at multiple scales and it is often possible to label regions of an image that contain very little visual information, if seen in the appropriate context. For example, we can identify a pupil or a nostril in a photograph of a face, even when the images these features produce are highly generic and would not be recognizable in isolation, even without blurring (Fig. 4). Again, it is clear that

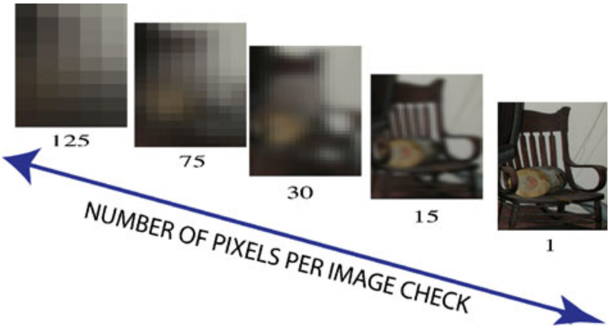

Fig. 3 Stimulus example from [9]. The number of pixels per image check provides a measure of information in the image

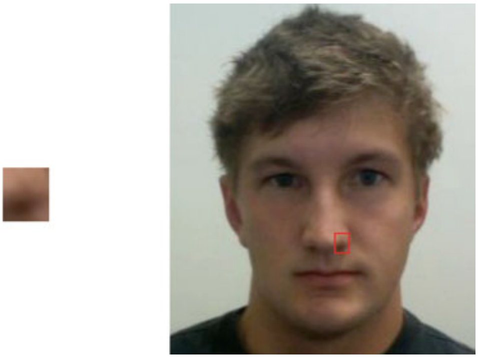

Fig. 4 Context aids in the recognition of non-degraded stimuli as well. The small image on the *left* is difficult to identify as a nostril outside of the context on the *right*

the human visual system must solve the local problem based on the global context. Thus, it appears that context-based recognition is most likely the norm, rather the exception.

2.4 Scene and object recognition: forest before the trees?

We have defined context as *any* information not due to the appearance of a target objects, including other objects in the scene as well as non-visual information. However, the global scene (e.g., a kitchen) in which an object appears may play a special role in contextual facilitation. While scenes are ultimately made up of collections of objects, some researchers have suggested that scene analysis operates before, and independently of, individual object recognition, based on global properties of images [45, 76, 98]. In this case, identifying the category of a scene (e.g., a kitchen or a mountaintop) might precede object recognition and serve to facilitate it. There are a number of sources of evidence for specialized scene recognition. First, people are able to extract the conceptual ‘gist’ of a scene even under very brief presentations [32, 89, 90, 109] and from a blurred image [98]. In addition, certain brain regions appear to be selective for places, rather than things (e.g., the Parahippocampal ’place area’) [29, 30]. There is also some evidence that people can extract certain truly global properties from an image such as the mean size [4, 19, 20] and center of mass [1] of simple stimuli such as circles. However, currently it is unclear whether scene categorization depends on global information rather than identifying critical objects. One important thing to note is that human observers do not typically face the challenge of determining a scene category very frequently. In everyday life, the category of the scene in which one is embedded is typically highly stable and also highly predictable from past experience (e.g., entering through your office door will not likely yield a farm scene). Thus, while it is likely that the contextual scene influences object detection and identification, it is not certain that scene perception itself is a critical part of this equation.

2.5 Context modeling, attention, and saliency

Another important potential role for context is in the allocation of visual attention, which is necessary in order to overcome the ‘overload’ of information available in a scene. Models of human behavior generally assume that there are two basic mechanisms driving the allocation of visual attention (typically accompanied by eye movements). ‘Exogenous’ attention is bottom-up, driven by features of the stimulus itself. This includes ‘involuntary’ orienting to abrupt changes in the image, such as the introduction of a new object [127, 128]. It can also include ‘voluntary’ orienting to regions of high contrast with regard to some ‘low-level’ property of the image such as luminance, color or orientation [118], leading to the notion of a ‘saliency map’ [64, 67]. ‘Endogenous’ attention, on the other hand, is based on top-down mechanisms such as previous expectations—e.g., the belief that important information will be present in a location—or goals of the observer, such as following instructions to attend to a specific region [88]. Endogenous attentional shifts may depend on higher-level properties, such as shifting attention between the people in a room. The role of contextual information in attentional allocation seems to fall somewhere in the middle of these two models of attention. On one hand, employing context often relies on a higher-level of processing—for example, incorporating relational information [22–24, 42, 43] or identifying the objects in a scene [14, 15]—which is typically not incorporated in models of exogenous attention. On the other hand, unlike standard endogenous attention, context may influence attention even without explicit awareness of the information [14, 15, 22–24, 42, 43].

Recent studies of attentional allocation while viewing familiar scenes present a complex interaction between top-down and bottom-up processes [81]. Torralba [111, 116] proposes a model of contextual cueing for attention guidance based on global scene identification and local saliency. In this model, the input image is analyzed in two parallel pathways: the *local* pathway computes typical image saliency and can be used to perform object recognition on the basis of local appearance. The *global* pathway computes global image statistics in order to identify the scene category, which serves to predict the presence or absence of objects as well as to predict their location, scale, and appearance *before* exploring the image. This in turn is used to modulate the allocation of the local pathway, based on prior probabilities associated with the identified scene.

2.6 Summary

Overall, there is a great deal of evidence that context facilitates object recognition in the human visual system. While the precise nature of this facilitation remains somewhat murky, a number of lessons may be drawn from this research with implications for computer vision. First, there is no doubt that context can provide information about the likelihood of specific objects being present in a scene (even according to the ‘response bias’ interpretation of contextual facilitation of Henderson and colleagues). In addition, it is clear that context can serve to disambiguate degraded images that cannot be recognized in isolation. In short, context provides critical *information* which can serve to supplement the information available from the object itself. While the precise scope of contextual facilitation in human vision remains somewhat controversial, there is incontrovertible evidence that context can, in some cases, provide critical information for object identification. Since this is an indisputable aspect of human visual perception, it should be modeled and incorporated into computer vision solutions.

3 Foundations of contextual modeling in computer vision

3.1 Early work

The realization that context plays a crucial role in human visual processes such as scene understanding and object recognition has motivated computer vision researchers to attempt to model and emulate such knowledge and behavior in computer vision systems and solutions.

The work by Yakimovsky and Feldman [124] is probably the earliest reference in the computer vision literature in which the authors employ contextual information to solve a classical image processing problem, in that case, image segmentation. Yakimovsky and Feldman present a "theoretical framework for a general system incorporating context dependence in a region analyzer" [124]. Their framework employs Bayesian decision theory techniques and uses problem-dependent information (semantics) to solve the image segmentation problem. Their goal is to obtain a partition of the input image and interpretation for the segments (regions) and boundaries so as to maximize the likelihood of having the right interpretation, e.g., that a segment interpreted (labeled) as 'sky' is above another one named 'hill'.

The next wave of early references to the use of context in computer vision can be found in the work of Strat, Fischler and colleagues [37, 103, 105–108] and a few others (e.g., [75]) in the early 1990s.

These early efforts usually consisted of hand-engineered, pre-defined, if-then rules, which attempted to emulate common expert knowledge in a narrow domain. However, these earlier methods were limited in their ability to deal with the uncertainty of real world scenes, which became the main focus of more recent systems, which typically rely on statistical models that are fit to data, as we shall see later in this paper.

3.2 Types of context

In this subsection we present a summary of different attempts to organize the interpretations and types of context into meaningful groups and categories available in the literature. There is no universal agreement or taxonomy on this topic. What follows are some representative examples of classification of types of context (and associated contextual modeling techniques).

3.2.1 Semantic, spatial, and scale contexts

Galleguillos and Belongie [39] refer to three main types of contextual information that can be exploited in computer vision solutions:

- Probability (semantic) context: refers to the likelihood of an object being found in some scenes but not in others. From the point of view of modeling, the semantic context of an object can be expressed in terms of its probability of co-occurrence with other objects and its probability of occurrence in certain scenes.
- Position (spatial) context: corresponds to the likelihood of finding an object in some positions and not others with respect to other objects in the scene.
- Size (scale) context: exploits the fact that objects have a limited set of size relations with other objects in the scene.

From a computational modeling viewpoint, Galleguillos and Belongie [39] observe that scale context might be the hardest relation to access, because it requires a more detailed information about the objects in the scene, consisting of the identification of at least one other object in the setting as well as the processing of spatial and depth relations between the target object and other object(s). They also claim that semantic context is implicitly present in the other two types of context—spatial context and scale context—, although it can be obtained from a wide variety of other sources, such as strongly labeled training data and external knowledge bases.

3.2.2 Things and stuff

An alternative terminology was proposed by Heitz and Koller who introduced a "Things and Stuff" (TAS) context model [51]. In their work, the terms 'stuff' and 'things' (originally introduced by Forsyth et al. [38]) are used to distinguish "material that is defined by a homogeneous or repetitive pattern of fine-scale properties, but has no specific or distinctive spatial extent or shape" (*stuff*) from "objects with specific size and shape" (*things*). Heitz and Koller claim that "classifiers for both things or stuff can benefit from the proper use of contextual cues". Consequently, they present four possible categories of context-modeling techniques:

- Scene-Thing context: refers to models which allow "scene-level information, such as scale or 'gist', to determine location priors for objects".
- Stuff-Stuff context: captures notions such as "sky occurs above sea" and "road is likely to appear below building".
- Thing-Thing context: considers models that take into account the co-occurrence of objects, and encode, for example, that a tennis racket is more likely to co-occur with a tennis ball than with a lemon (see [93]).
- Stuff-Thing context: enables texture regions within the scene to add predictive power to the detection of objects present in the scene.

3.2.3 SBC and OBC

Rabinovich and Belongie [92] have proposed a classification of contextual models for computer vision (in general) and object recognition (in particular), consisting of: models with contextual inference based on the statistical summary of the scene (which they refer to as Scene Based Context models—SBC) and models representing the context in terms of relationships among objects in the image (Object Based Context—OBC). When comparing their classification terminology with that of Heitz and Koller [51], they underscore the fact that "neither the SBC nor the OBC models explicitly separate thing from stuff" and defer the formulation of SBC and OBC models in terms of a "thing vs. stuff formalism" to a future time, when the "thing vs. stuff distinction becomes more rigorous".

3.2.4 A broader view of context

Divvala et al. [27] start from the definition of context as "any and all information that may influence the way a scene and the objects within it are perceived" [104]. They compile a list of the many different sources of context that have been discussed in the literature, and add some of their own, resulting in a broader and longer list, as follows:

- Local pixel: “captures the basic notion that image pixels/patches around the region of interest carry useful information” [27]. Examples of local pixel context include: image segmentation, object boundary extraction, and several object shape/contour models.
- 2D scene gist: refers to models that use global statistics of an image to capture the “gist” of a scene (e.g., [79]).
- 3D geometric: corresponds to models that attempt to capture the coarse 3D geometric structure of a scene, or its ‘surface layout’ (e.g., [58]), which can then be used to reason about supporting surfaces, occlusions, and contact points.
- Semantic: used to indicate the kind of event, activity, or other scene category being depicted as well as the presence and location (spatial context) of other objects and materials in the scene.
- Photogrammetric: related to several aspects of the image acquisition process, including intrinsic camera parameters (e.g., focal length and lens distortion) as well as extrinsic ones (e.g., camera height and orientation).
- Illumination: “captures various parameters of scene illumination, such as sun direction, cloud cover, and shadow contrast” [27].
- Weather: used to “describe meteorological conditions such as current/recent precipitation, wind speed/direction, temperature, season as well as conditions of fog and haze” [27].
- Geographic: refers to information about the actual location of the image (e.g., GPS coordinates), or more generic information such as terrain type (e.g., tundra, desert, ocean), land use category (e.g., urban, agricultural), elevation, and population density, among others.
- Temporal: captures “temporally proximal information, such as time of capture, nearby frames of a video (optical flow), images captured right before/after the given image, or video data from similar scenes” [27].
- Cultural: refers to biases and intentions involved in the processes of taking pictures (e.g., framing, focus, subject matter) and selecting datasets, among others.

The list above includes contextual aspects that can be captured from other sources, i.e., it reaches far beyond what can be captured from visual input alone.

4 Spatial knowledge and contextual reasoning

A common point of agreement between all previous taxonomies is the prominent role of spatial knowledge in contextual reasoning. All the key entities in a 3D physical scene are localized (e.g., objects, cameras, light sources) and this assertion is consequently also correct in the image space. Each object of interest in a given image is supposed to fit into reasonable spatial relationships with other related objects or, more generally, with the overall indoor or outdoor environment. Notwithstanding that many definitions of context coexist, the context primarily gives access to these spatial relationships and prior knowledge.

All the spatial entities playing a role in computer vision may be involved in these spatial interactions. This is true both in the image domain (e.g., pixels, edges, patches, superpixels, regions, segments and their related descriptors) and in the geometric domain (points, lines, surfaces, volumes and their algebraic representations).

Interpixel relations are extensively used with pixel attributes ranging from color [36] to appearance descriptor [121] or semantic labels [69].

Pairwise relations between patches [101], regions, or segments [41] are at least as important as statistics between pixels. There is indeed a strong tendency to consider regions as the basic contextual entity, since the spatial extension of a physical object in image space is naturally an image region. The spatial extent of the regions may depend upon the use case: bounding boxes for object localization, patches for early part detection, object-shaped regions for *things* segmentation, amorphous segments for *stuff* representation.

Global image features are also crucial for top-down reasoning about the overall scene appearing in an image. The 'gist' [113] is perhaps the most famous example of global statistics that capture a discriminative summary of the whole image by applying filter-banks. Whilst the gist is 2D, a coarse 3D layout of the scene [49] is also an alternative spatial entity to constrain scene-object relations.

In order to reason qualitatively about the 3D relation between the scene and a camera, properties such as vanishing points, parallel lines [49], and horizon line [59] can be detected. Furthermore, qualitative reconstructions of the environment from a single image can be based on surfaces [58] or 3D blocks [47].

Of course, the interactions between all these spatial entities ultimately have an echo in the semantic domain by establishing relations between objects or object parts. In the next section we review the major spatial relations and priors effectively used in context modeling and contextual reasoning.

4.1 Spatial relations

4.1.1 Occurrence

Co-occurrence is one of the simplest approaches to introduce relationships between objects in a visual scene [17, 41, 93, 120]. Contextual interactions such as "cars appear on roads" can be translated directly in contextual relations between object labels. In this case, the presence of a certain object class in an image (e.g., a road) statistically influences the presence of a target object (e.g., cars). It is straightforward to build context matrices to count co-occurrence of labels given a dataset where many objects are labeled. Such co-occurrence matrices can easily be translation- and rotation-invariant and regularly show robustness to affine and pose changes [120]. It is also well known that certain objects (e.g., computer monitors, beds) occur more frequently in some places (e.g., offices and bedrooms, respectively) [113]. Starting from these learned co-occurrence statistics, Rabinovich et al. [93] devised interaction potentials for Condition Random Field (CRF) in order to measure contextual agreement between detected objects. It is interesting to notice that the terms "semantic context" and "co-occurence" are sometimes used interchangeably [41]. The statistical model proposed by Carbonetto, Freitas and Barnard [17] also learns co-occurrence between concepts (e.g., image caption words). However in their model, Markov Random Field (MRF) interaction potentials are estimated only between neighboring image segments (e.g., object blobs). More so than co-occurrence, such potentials also describe the ‘next to’ relationship between object labels. Wang et al. [121] give a formal definition of co-occurence and occurrence functions. These functions provide a probability distribution of labels over different regions centered around a given labeled pixel. Perhaps the most interesting idea in this scheme is to relate two independent sets of labels by using occurrence (and not only co-occurence). This integrates both shape and appearance labels into a shape and appearance context descriptor.

The occurrence analysis can be also made between object parts (e.g., detecting nose and mouth as part of a face). In this case again, the relative position of the parts is crucial [33]. Fink and Perona [36] detect faces by using both the individual detections of M parts/entities (left eye, right eye, mouth, nose, entire face) and their spatial arrangements. Hence, they treat M entities at each boosting iteration and compute M maps that give the likelihood of each entity appearing in different positions in the image. Combining face parts is made possible by using all the likelihood maps as additional input channels for subsequent boosting iterations. Consequently, the likelihood map for eye detection can be used to further detect mouth, and likelihood maps for face detection can be helpful to detect multiple faces, since faces tend to be horizontally aligned in the considered data set. Practically, the famous rectangle contrast features (from Viola and Jones [119]) are used as weak learners to select relevant arrangements of spatial entities and filter out non-relevant ones. A large contextual window is used to analyze such co-occurrence and spatial relations. Another similar approach is brought by Perko and Leonardis [87], in which horizontal alignments between pedestrians in street scenes are learned by estimating a 2D probability distribution of other pedestrian locations given a pedestrian in the center of the image.

4.1.2 Spatial arrangements

As previously suggested, the relations between spatial locations of objects lead to additional arrangement constraints (“car appears *on* roads”) and go beyond the co-occurrence relationships. Spatial expressions in natural languages are a good way to describe such relations that often turn out to be rather qualitative (e.g., “object A is *around* object B”) rather than quantitative (or metric). These relations can be expressed and used both in 2D (image space) or in 3D (object space).

We first distinguish three classes of qualitative 2D relations:

(i) direction relations;
(ii) distance relations; and
(iii) topological relations.

Direction relations (i) express the direction of one object (the primary object) relative to another (the reference object). Such relations can be defined if a frame of reference is known. Usually the cardinal directions (E,N,S,W) and their refinements (NE, NW,SW, SE) can be used (tacitly) as an extrinsic frame of reference. One can also assume an intrinsic orientation of reference in the image space (so that we can talk, for example, of an object being to the “left” of a building). The relative vertical positions (“above”, “below”) are frequently used and judged discriminative enough to detect object in conventional dataset like PASCAL where as horizontal positions do not necessary carry much discriminative information.

Distance relations (ii) can be divided into two major categories: those which provide measurements on some absolute scale and those which provide relative measurements. In normalized image spaces some distances can be expressed in terms of (absolute) pixels (e.g., "object A is about 200 pixels away from object B"). Heitz and Koller [51] also cite some human knowledge e.g., "cars park 20 feet away from buildings" that highlight the limitation of 2D spatial reasoning with a single image—since a 3D (geo)metric context would be required to capture that relation. More frequently, relative measurements lead to 2D qualitative relations such as "close", "far", or "equidistant". Determining the correct scale and associated thresholds for such relations is a difficult task in the general case, yet tractable in domain-specific applications.

Topological relations (iii) describe the relationship between an object and its neighbors. Formally defined by considering interior, boundary and exterior of an object, intersection relations such as "touches", "overlaps" , "contains" (in, inside), and "crosses" are often used in practice. Some authors also propose a slightly nuanced version ("encloses") of the simpler relation "contains" [69, 101]. One should also notice that the simplest topological relation is when two regions/objects are "disjoint".

Of course, all these spatial arrangements can be further combined. For instance Singhal et al. [101] use "far above" and "far below" to mix direction and distance relations. They also mingle "left" and "right" relations and introduce a weaker "beside" relation. In their TAS model, Heitz and Koller [51] also combine all types of relation (eight directional relations, two different distances, and a topological "in" relation) to generate many candidate relationships (25, to be exact) from which they extract the most useful ones.

As mentioned earlier, the main weakness of 2D qualitative spatial arrangements lies in the fact that the correct scale of the relationships must be detected or learned *a priori*. Recent works on "geometric context" [47, 50, 97] that attempt to infer qualitative 3D reconstruction from a single image may provide a solution. If a 3D context is recovered, the appropriate scale may be easier to determine. In addition, 3D relationships can be also stated between volumes: 3D pairwise depth relations [47] such as "in front" and "behind" have been used along with 3D support relations such as "supports", "(is) supported by". The relations "front"/"behind" can be interpreted as *order* relations.

It is interesting to note that order relationships involving the *size* of the considered objects/regions are rarely cited in the literature. For instance, relations such as "larger"/"smaller" or "greater than/less" are not frequently used in contextual object detection or scene understanding. This may reflect the fact that true size (as opposed to apparent size) can only be determined based on a recovering 3D depth information.

4.1.3 The discriminative alternative

From the computational point of view, there are multiple possible representations of the previously discussed types of relations by generative models or processes. Carbonetto et al. [17] and Galleguillos et al. [41] use interaction potentials for MRF and CRF respectively, Heitz and Koller [51] activate or not each relation by using indicator variables in their TAS model, Gupta et al. [47] visualize their relations in a 3D parse graph. In these approaches, one might say that the spatial relations are handled explicitly.

These approaches, although promising, can be challenged by the alternative discriminative approach where the spatial relations are treated implicitly by classifiers. Both co-occurrences and spatial arrangements can be learned by using sampling patterns [87, 120–122]. The first step in such discriminative approaches is to compute many features at each pixel location during the learning stage. The basic assumption is that contextual information can be stored as layered images or feature maps. As an example, Wolf and Bileschi [122] consider a 20 dimensional feature vector by concatenating color and texture features, semantic labels and relative position measurements, at each pixel location. More recently Perko et al. [87] also integrate geometric feature maps [57] and saliency maps. At this stage an initial natural image is converted into an image containing many layers of information.

The second step consists of extracting a context descriptor given a candidate object location. A sampling pattern can be designed to collect the contextual layered informations at predefined locations centered around the candidate object. Biologically inspired, polar sampling patterns [87, 120, 122] are popular. For instance, Wolf and Bileschi [122] use 40 relative polar locations (multiple radii and multiple orientations) and then concatenate their 20 dimensional features to finally output a $40 \times 20 = 800$ dimensional context descriptor.

In support of techniques working at the descriptor level, Zheng and colleagues [120] argue that such polar sampling offers greater flexibility in capturing different types of context (including *thing-thing*, *thing-stuff*, etc.) and in representing many existing spatial relations (including “inside”, “outside”, “left”, etc.). Since feature vectors naturally covary between the sampling locations, such context descriptors, fed into well-chosen classifiers, allow such systems to exhibit important correlations and discriminant signals for object detection. Simultaneously, irrelevant relations and unhelpful signals with respect to the visual task are automatically ignored by learned classifiers.

These advantages probably explain why the work by Wolf and Bileschi [122] has been followed up by several other authors: Wang et al. [121] have introduced a sampling pattern made of concentric square rings, while Zheng et al. [120] have also proposed a polar geometric structure but use dense SIFT extractions in radial bins, whereas Perko et al. [87] have recently reused a multiscale version of the initial sampling pattern of Wolf and Bileschi [122].

4.2 Shape and location priors

In this section we discuss shape and location priors that complement the already rich spatial knowledge that can be used to recognize objects and understand visual scenes.

4.2.1 Shape priors

Intershape relations and shape priors are fundamental types of spatial knowledge for object detection, due to the fact that shape is perhaps the most important feature of an object. In order to improve object detection, many queries can be used: Is object A similar to object B? Do they have nearly the same shape? Is object A a part of object B? Does an object seem like a chair, a table, a plane? Therefore many authors have studied and still use shape context [10, 121]. We review three main approaches for modeling shape priors at both object and scene levels.

Firstly, many recent papers [35, 72, 82–84] propose methods for learning object shape models and shape alphabets by analyzing the arrangement of edge features such as the pairwise interactions between edge features or the relative positions of edge features with respect to the centroid of the shape. As an example Ferrari et al. [35] present a family of scale-invariant local shape features formed by short chains of connected contour segments. Detecting faces by multiplying the window hypotheses over pixel locations at many scales works fine if the interior part of the target object is discriminant enough. Unlike faces, other objects category like the ones handled by Ferrari et al. can be more easily detected by using their boundaries.

Second, learning deformable shape models for part-based detector is also becoming very popular. Felzenszwalb and colleagues [33] use star-structured deformable part models (e.g., shape grammars) and associated cascade of part detectors. Yang et al. [125] define a part-based model of shape prior hierarchically. At the elementary level, the simplest shape prior is a soft segmentation mask (or alpha-matte) which records the probability of a pixel belonging to the object at some location relative to the detection center. Then shape priors are further refined as a mixture of parts models and depend of the object pose (e.g., side vs. frontal cars). Shape priors have indeed a fundamental role when tackling interleaved object recognition and segmentation. Obj Cut [68] is perhaps one of the most influential works in this area: the key idea is to introduce a part-based model as a prior knowledge of object shape to supervise grouping-based segmentation.

As our third step, we finally put the emphasis on geometric context viewed as a global shape prior at scene level. The influential work of Hoiem et al. on geometric context [57] introduces a rough 3D sense of scene geometry as a key contextual component. They consider outdoor images and define three semantic classes associated with the ground, the sky and the vertical entities like buildings, trees etc. This typology of outdoor surface classes can be viewed as the main prior knowledge in their statistical learning approach. The major idea is to map stable image segments to the main planar surfaces of an outdoor environment. Practically each superpixel [34] of a single image is classified leading to three likelihood maps, one for each class (e.g., ground, sky, and vertical surfaces such as buildings). Additionally, vertical planar surface elements are further labeled with their 3D orientation. Texture- and edge-based features are used to provide such orientation cues. The interest of such a work is at least threefold. First, the labeled images can be directly used as contextual features [87]. Second, the labeled images can be "popped-up" to qualitatively reconstruct a coarse scaled 3D model of the scene from a single image. To perform such a 3D elevation, a key problem is to cut and fold the labeled image properly and this explains why occlusion reasoning [60] has been extensively studied by the same team. Third, a rough 3D scene geometry can interplay with low level object detectors [59]. In this particular work [59], Hoiem and colleagues make a step forward and infer simultaneously object labels (e.g., pedestrian or car detections), qualitative 3D geometry (e.g., labeling and orientation for sky, ground, vertical surfaces in street scenes) and camera viewpoint. A coarse 3D estimation of the scene geometry along with an approximate camera viewpoint naturally lead to prior knowledge about where a pedestrian or a car might be both in 3D world and in the 2D image space.

Hedau, Hoiem and Forsyth [50] address the case of indoor scenes with the same approach based on the fitting of a coarse 3D geometric model to room images. The shape prior for the room's geometry is basically composed of perpendicular surfaces (labeled as "walls" and "floor") leading to a 3D box layout. Then, object detections are performed "inside the box". Reasoning again in the 3D space, the sliding window strategy for object detection is generalized as a sliding 3D cuboid procedure. Therefore beds (frequently occurring in rooms) can be detected and located as 3D parallelepipeds by integrating both appearance and geometric coherence. Outdoor scenes can be treated as well by introducing the cuboid shape prior for building detection [47].

4.2.2 Location priors

These last cited works suggest the importance of location priors in contextual reasoning. Without paraphrasing previous parts of our discussion, we now briefly recall to the reader a few solutions to represent and encode object location priors in computational models. First of all, it is well known that Torralba's 'gist' can be used to predict the vertical location of object classes. This is called *location priming* [113]. Less influenced by human perception aspects, Wolf and Bileschi [122] propose a much simpler position descriptor that allows a simple classifier to learn a wide variety of position priors. In their context descriptor, Wolf and Bileschi [122] simply employ ten features to calculate at each pixel, namely the distance to ten predefined positions spread over the image. They report good results with this straightforward position prior applied to street scenes.

Information about the camera's location is also a key component of the geographic context [27, 48]. The geolocation of the camera can naturally be used to infer prior knowledge about the observed scene. The tagged images from the internet and associated keywords are booming sources of contextual informations for some real life applications like automatic image annotation. As an additional significant example, the third eye (e.g., contextual satellite images) [73] also provides impressive power of understanding only given the GPS coordinates (e.g., metadata) attached to an image. The role of the camera position is finally fundamental in determining the actual object detection scale depending on the distance between the observer and the object. Scale selection is a really difficult problem in image understanding. Qualitative 3D reasoning has a great potential in this domain: Hoiem et al. [59] manage to select detection scale for pedestrians or cars by introducing a prior on camera height (1.67 m as an average eye level for an adult) and on horizon line position.

4.3 Contextual reasoning

Contextual reasoning consists of integrating appearance processing and the aforementioned spatial relations or prior knowledges to solve a rich set of visual tasks. Among them, we would like to cite object presence and counting problems along with position and scale detection subproblems, image segmentation, scene layout approximation, depth estimation and qualitative reconstruction, and semantic interpretation.

4.3.1 Principled integration of context and appearance

We first discuss basic strategies to integrate appearance and contextual information while focusing on object detection applications.

In Section 4.1.1 we discussed the work of Rabinovich et al. [93] that uses co-occurence context to reduce ambiguity in object appearance by using other objects in presence. The idea behind this contextual association is to define context as a function of previously recognized objects. As a consequence, a possible architecture to combine appearance and context is to first build multiple detectors: some for the related objects, others for the target objects. The outputs of the related object detectors then help to visually detect target objects: for instance they may help to infer their likely presence and/or likely location. The limiting factor in this approach is that estimating the context can be as arduous as detecting each related object. In their critical survey, Wolf and Bileschi recommend avoiding this approach. According to their domain-specific—limited to street scenes—experiments, accurate context can be determined from low-level early visual features [122].

Another very simple way to combine contextual cues and appearance signals is to put them together in a feature vector. As an example of such an approach, Wolf and Bileschi [122] simply concatenate semantic labels (boolean presence variables indicate if each pixel is over a building, a road, the sky etc.) and low level appearance information before training classifiers in a purely discriminative framework (see Section 4.1.3).

More generally, context can be used as a post-filtering or pre-filtering component in an object detection procedure. On one hand, numerous modeling attempts [93, 101, 120] use the context as a post-processing technique: the candidate objects are first detected by an appearance-based approach and then false positives are filtered out by using context information. In this case, a low threshold must be associated with the appearance detector to ensure that most of the true positives are passed on to the second stage. On the other hand, Wolf and Bileschi [122] propose to use a rejection cascade to combine appearance and context cues. In one version of their work a context-based descriptor is first used and acts as a preliminary filter before an appearance detector. Pixels are passed to the appearance detector if and only if they are classified as plausible according to the context. This is one of the rare examples where context is used as a pre-filtering technique.

4.3.2 Probabilistic fusion

In order to combine appearance and contextual cues in a mathematically sophisticated way, many probabilistic models have been proposed. Generally speaking, most of them employ a MAP-like approach by maximizing, at test time, a posterior confidence score which combines a prior detection score with a contextual confidence score given a candidate region/window for detecting an object. This approach usually comes along with the aforementioned post-processing approach (first detect candidate objects with appearance-based descriptors and then filter out the false positives contextually). In other words, the prior detection score is an output of the recognition system and the second term introduces semantic structure and/or contextual constraints.

Generally speaking a probabilistic fusion model attempts, *at test time*, to maximize a general criterion that can be expressed as:

$$\text{posterior confidence score} \propto \text{prior detection score} \times \text{contextual confidence score} \tag{1}$$

We first discuss the prior detection score term. Very often it comes from a common sliding window object detector. Mostly cited detectors are publicly available (HOG, UocTTI, boosting approaches). Such discriminative detectors generally output a conditional probability that a candidate window W_i extracted from an image I contains the ith target object labeled as O_i: $P(O_i|W_i)$. In their empirical comparison of multiple context representations, Divvala et al. [27] explicitly decompose the detection score assigned to an image I in three terms related to object presence $P(O_i|I)$, location $P(x_i|O_i, I)$ and size $P(h_i|x_i, O_i, I)$, respectively. This distinction is very interesting because the *object presence* probability can be generalized for multiple instances of the same object class. Torralba et al. [113] call this problem *object counting*. The associated probability can be written $P(O_i^N|I)$ where $N \in \{0, 1, 2, 3\text{–}5, 5\text{–}10, >10\}$. When multiple object classes can be independently detected the detection score is simply a product $\prod_i P(O_i|W_i)$. In their maximum margin context model (MMC), Zheng et al. [120] also introduce an importance factor α such that the prior detection score $= P(O_i|W_i)^\alpha$. For $\alpha > 0$, the larger is the importance factor, the more important the prior detection score is and the lower contextual confidence tends to be. It should be noted that in some cases the outputs from a recognition algorithm are not probabilities, e.g. in the case of SVM outputs. A logistic regression function [51] can then be fitted to map a margin score to the desired domain. A more empirical normalization using robust statistics is also recommended by Perko et al. [87].

The second term in the general equation (1) varies considerably among different approaches in the literature. We review some important types of contextual confidence score.

Rabinovich et al. [93], mainly study co-occurrences between objects (e.g., contextual agreement between the labels assigned to image segments). In a CRF framework, interaction potentials ϕ between segment labels form their contextual confidence term proportional to $\exp\left(\sum_{i,j} \phi(O_i, O_j)\right)$. Galleguillos et al. [41] do the same except that the interaction potentials ϕ_k are redefined for each of the four pairwise spatial relationships k they consider.

In their probabilistic *Things And Stuff* (TAS) model, Heitz and Koller build a context by linking the detection of objects (e.g., ‘cars’) with the presence of unsupervised image segments (e.g., automatically detected stuff clusters representing ‘road’ appearances). For that purpose, they use a set of indicator variables $R_{i,j,k}$ to indicate whether the detection of object O_i in presence of a stuff/contextual segment S_j are related by a likely spatial relationship k (e.g., the kth relation might be the ‘above’ relation). Hence their contextual confidence score first tries to maximize a conditional probabibilty $\prod_{j,k} P(R_{ijk}|O_i, S_j)$ and simultaneously (by independence) a joint probability (e.g., a generative model) to classify each extracted image feature F_j as a stuff segment S_j: $\prod_j P(S_j, F_j)$.

Torralba and colleagues [113] use the scene ‘gist’ g as a global image feature to build a twofold probabilistic contextual confidence score. The gist provides ‘top-down’ information that allows one to predict: 1) how many object instances should be present; and 2) where they might be located. The prior detection score c_i is first converted in a local gaussian likelihood $P(c_i|O_i)$ and combined in closed form with a another local gaussian likelihood term which represents the expected location x_i given the candidate object class and the gist $P(x_i|O_i, g)$. Then presence and counting priors are further integrated in two steps. The type of scene (e.g., street, highway, forest etc.) is first predicted from the gist as done by Quattoni and Torralba [91]). And, given the scene category, the plausible number of object instances within each considered object class is then integrated in the contextual confidence score.

As previously stated, Hoiem et al. [59] detect objects (e.g., presence O_i, location x_i, and size h_i for the i*th* object) while estimating a qualitative 3D geometry g_i around each object [57] and a camera viewpoint C. They first integrate the prior detection score, the mutual dependency between the object scale and the viewpoint and a prior on viewpoint: $P(O_i, x_i|I)P(h_i|CI)P(C)$. The two last terms participate in the contextual confidence evaluation. Additionally a local geometry evidence $P(gi|I)$ and the geometric consistency between the object surface and nearby surfaces $P(g_i|O_i)$ are properly associated to complete the contextual term. Similar models integrating geometry estimation and object detection have been recently proposed for indoor scenes [50] and outdoor scenes [47]. In both cases 3D shape priors are provided to detect prominent objects in the scene by simultaneous integration of appearance and geometric hints.

In their MCC model, Zheng et al. [120] use a very innovative 'context positiveness function' to serve as a contextual confidence score. This function explicitly measures the risk of using contextual information tackling the issue of decision-making for contextual reasoning.

Beyond independence assumptions, Perko and Leonardis [87] compare several ways to combine appearance and contextual features. They first assume that prior detection scores and context confidence scores are statistically independent and they simply fuse them by multiplications (1). In their setting, the detection performance is decreased by such a naive fusion. It turns out that all their contextual cues (geometry, texture, saliency, co-occurrence) are not equally relevant and that the independence assumption is obviously false. For instance saliency maps, used to contextually focus on salient regions when searching for an object, are useless and contaminate the fusion. In their subsequent test they explicitly model the dependencies between the appearance scores and the contextual confidence scores using a kernel density estimation (KDE) for the associated joint densities. They report better results in that case, empirically showing that local appearance and contextual informations are non independent.

In our description and review of selected probabilistic models to combine appearance and contextual informations we have focused, for readability reasons, on assessment criteria. In their recent survey paper on this topic, Galleguillos and Belongie [39] take another route—complementary to ours—and present two main classes of machine learning techniques (e.g., classifiers and graphical models) as the most popular way to learn the probabilistic relation between appearance and context.

4.4 Summary and implications

In this section, we have shown that state-of-the-art context representations are mostly related to spatial knowledge. We have kept the discussion within the domain of still-image understanding and intentionally did not discuss the temporal dimension and associated cues. Within our scope of investigation, we do believe that spatial relations, along with shape and location priors are prominent in computational aspects. In some sense, the critical survey of Galleguillos et al. [39] confirms our claim, since they conclude that *spatial and scale context involve using all forms of contextual information in the scene* and that *semantic context is implicitly present in spatial context*. If we admit that the formal representation of spatial arrangements and relations to priors is the key, we further think that decision taking about context will be a major challenge for future works. The first idea is that context is efficient when appearance is weak. So before using context we have to formally detect when signals are degraded and when appearance is deficient. Selecting the adapted strength of contextual constraints is linked with assessing appearance expressiveness. This issue is still open in our opinion. Furthermore, recent papers that we reviewed [41, 51] highlight the need for automatically learning the best and most discriminative contextual relations to be used. Combining spatial relations sequentially or globally for contextual reasoning is probably limited. A better way might be to select active and useful relationships given a context performance metric. Wolf and Bileschi [122] measure the context performance by quantifying how much aid is given to subsequent stages in the detection process. Rabinovich et al. formally define both confidence and amibiguty of a segment labeling by considering a distance between the best labeling and the second best given a segment [93]. More sophisticated definitions of visual ambiguity might be a necessary ingredient of real context performance metrics. In this domain, the MCC model by Zheng et al. [120] is interesting since it is basically built to reduce such an ambiguity criterion. Finally the integration of multiple sources of context will lead to many papers in the upcoming years. The holistic scene understanding problem will be intensively revisited by integrating recent progresses in object detection, image segmentation and scene reconstruction.

5 Implications and applications

5.1 Implications

The understanding of how visual processes and tasks such as visual search, object detection, recognition, classification, and autonomous navigation—among many others—are strongly influenced by contextual cues is crucial to the advancement of the state of the art in both human and computer vision research. Such understanding may lead to better, i.e., more robust and reliable, computational models of the contextual influences caused by a quick understanding of the scene (after having captured its gist) and/or mutual relationships among objects in the scene.

As the field of contextual modeling matures, researchers will have to start working on answers to relevant associated questions, such as:

- How can we measure success in context modeling? While learning about and incorporating contextual information in computer models provides a distinct set of hurdles and challenges, the goal of context modeling is ultimately to improve performance in a functionally useful application. Thus, simply demonstrating that contexts can be learned and/or modeled does not provide a good metric

for the utility of such a model. Nevertheless, it may be necessary to consider the inherent strengths of a system *generically*, without regard to a specific application, since it is likely that a good model will be useful for multiple applications.
- What is the appropriate input of a context modeling algorithm? In theory the input could consist of raw images, segmented images and/or labeled images and could also include metadata such as keywords or tags. In addition, a model could use the image data in order to extract a scene category—as in Torralba's model. Alternatively, in some circumstances the scene category could be provided beforehand as a form of metadata as well. For many applications (e.g., robotics), the environment in which a system is implemented is often highly constrained and predictable and it might not be necessary to perform scene analysis independently prior to implementing the context model. However, other applications, such as web-based image search, might require scene analysis at the front-end of the system.
- What is the appropriate level of scene category for modeling the context? Any given scene image may be categorized at many levels, including basic-level categories (e.g., outside/inside, home/office) mid-level categories (a children's bedroom; a doctor's office) as well as highly specific, unique categories (my own bedroom). The choice of scene category may well vary depending on the particular application and the environment in which it will be employed.

5.2 Applications

Contextual information plays a significant role in object detection, localization, and recognition tasks. In this subsection we look at other computer vision tasks and research areas that may benefit from contextual information, particularly:

- Semantic event recognition
- Visual search and retrieval
- Context-aware image annotation and photo management systems
- Autonomous navigation

5.2.1 Semantic event recognition

Luo and colleagues [65, 73] have proposed a system that combines visual cues from the picture with its geographical positioning systems (GPS) information, available as metadata. The longitude and latitude coordinates corresponding to the picture location are used to obtain satellite images of the environment in which the picture was taken, providing a "third eye" above the scene and its objects.

Their work made three significant contributions to the state of the art in semantic event recognition: (i) it launched a novel way to use satellite aerial images, through geotagging, to recognize a picture-taking environment from above (rather than at ground level); (ii) it demonstrated the effectiveness of a new vision algorithm that uses both structure and color features to characterize different environment categories, and multiclass AdaBoost to achieve reliable recognition in spite of the large variability (e.g., due to weather conditions) in the satellite images; and (iii) it combined satellite image-based recognition with classical vision-based ground image recognition into a robust integrated detection system in which each of the two views contributes to improved recognition capabilities.

5.2.2 Visual search and retrieval

There have been several attempts at using contextual information to improve the performance of content-based image retrieval (CBIR) systems. The introduction of the color correlogram descriptor (a color-based descriptor that takes into account the spatial relations of local properties) by Huang et al. [62] is often regarded as the earliest attempt in "context-based CBIR".

Examples of more recent work include Amores, Sebe, and Radeva [2], in which they proposed a context-based framework for medical image retrieval on the grounds of a global object context based on the mutual positions of local descriptors. Their approach was tested using intravascular ultrasound images, with better quantitative results than the baseline method by Huang et al. [62]. In [3], they have introduced a novel type of image representation—the Generalized Correlogram (GC)—and represented each visual object using a constellation of GCs, where each GC encodes information about some local part and the spatial relations from this part to others (i.e., the part's *context*). The proposed matching scheme exploits that representation and takes into consideration the spatial coherence between the matching of local parts. The GCs are spatially quantized using the log-polar spatial quantization originally proposed by Belongie et al. [10], which makes the correlogram more sensitive to local context.

Approaching context from a broader view, which includes tags, geospatial information, and other types of metadata has become a dominant topic in recent years. In a recent presentation,[1] Jain coined a new term (*contenxt*) to represent the confluence of content (data) and the context in which it is presented. Jain is among many who claim (and hope) that the modeling and use of context in visual information retrieval systems will help narrowing the infamous "semantic gap" problem [102] that permeates the field.

Other contemporary representative efforts that use context information—in a broad sense—to improve visual search and classification tasks include, among many others:

- *ContextSeer* [126], a system that re-ranks text-based image search results based on informative context cues that are automatically selected by the system;
- the work by Kennedy and Naaman [66], in which a combination of tags, visual features, and geolocation information is used to automatically select representative landmark pictures and discard non-representative ones.

5.2.3 Context-aware image annotation and photo management systems

There have been a number of very recent efforts that attempt to bring context awareness to image annotation and organization tasks. These efforts use the term "context" in a broader way and usually refer to contextual information gathered from sources external to the visual contents of the images, e.g., image tags, image file metadata, and GPS coordinates.

[1] http://www.slideshare.net/jain49/contenxt-100407

Cao et al. [16] have developed a system for annotation of collections of personal photos that exploits the contextual information naturally implied by each photo's associated GPS and time metadata. First, they employ a constrained clustering method to partition a photo collection into event-based subcollections. Subsequently, they use conditional random field (CRF) models to exploit the correlation between photos based on: (i) time-location constraints; and (ii) the relationship between collection-level annotation (i.e., events) and image-level annotation (i.e., scenes). The authors claim that by employing such a multilevel annotation hierarchy, their system addresses the problem of annotating consumer photo collections that requires a more hierarchical description of the customers' activities than do the simpler image annotation tasks.

O'Hare, Smeaton and colleagues [77, 78] have developed MediAssist, a system for browsing, searching and semi-automatic annotation of personal photos, which takes into account both image content and the context in which the photo is captured. This semi-automatic annotation includes annotation of the identity of people in photos based on a combination of context and content. It proposes language modeling and nearest neighbor approaches to context-based person identification, in addition to novel face color and image color content-based features (used alongside face recognition and body patch features).

5.2.4 Autonomous navigation

Siagian and Itti [100] have proposed a simple, biologically plausible, context-based scene recognition algorithm that captures the "gist" of a scene using a multiscale set of early-visual features, which are shared with a model of visual attention and encoded as a low-dimensional signature vector. The low-computational complexity of their approach makes it attractive to mobile robotics applications, e.g., classification of scenes and buildings in a campus environment, under which it has been successfully evaluated [100]. In the proposed system, sharing raw features between the 'gist feature extraction' block and the saliency model helps increase localization resolution by using salient cues to create distinct signatures of individual scenes and establish finer points of reference that may not be differentiable by gist alone.

6 Concluding remarks

In this paper we presented a survey of context modeling approaches in computer vision and some of their applications. All the previously discussed vision tasks and applications are notably arduous since they are intrinsically linked with the *inverse problem* that lies at the core of human and computer vision. A possible approach to overcome the inverse vision problem is to increase the sources of information used to recover real-world properties of a scene and make educated decisions about what they contain. The multiple contextual cues reviewed in this paper have the potential to play a significant role in this process. The emerging computational models of context recently proposed in the literature show great promise in enabling better solutions to many precise vision problems such as: object detection, object localization, and 3D structure estimation.

However, at the current stage of research, it is by no means settled what the correct model of context is and how it can best be leveraged in enhancing computer vision. Indeed, existing implementations of context modeling have only provided marginal improvements over previous technologies. Nonetheless, the enthusiasm with respect to this approach is not based on the success of early results but rather on the relative lack of success in computer vision using standard, non-contextual, techniques combined with the clear role contextual knowledge appears to play in human vision. Whether context turns out to be the 'missing link' in bridging the wide chasm in performance between artificial and biological vision remains to be seen. However, at the current moment it appears to be our best bet.

In the upcoming years, a *leitmotiv* within the computer vision community will probably be the holistic scene understanding problem which is, more than an inverse problem, a collection of inverse problems. Recovering the 3D layout of the scene, categorizing the scene, segmenting the image, recognizing the objects and identifying events are mutually dependent inverse subproblems which, combined, might interact to parse the semantic content of visual images.

Acknowledgements The authors would like to thank Geraldine Morin, Pierre Gurdjos, Viorica Patraucean, and Jerôme Guenard, for the insightful discussions and constructive suggestions.

Appendix: Resources for researchers in the field

In this appendix we provide a brief survey of some of the most prominent research groups working on context modeling in computer vision and associated topics, as well as associated resources, such as datasets and open-source code.

Research groups

Torralba, Oliva, et al.

The work by Torralba,[2] Oliva[3] and colleagues at MIT is among the most representative efforts in combining human behavioral and computational research on topics related to the broad themes of "visual scene understanding" and "object recognition".

The paper by Oliva and Torralba [79] in which they introduce the spatial envelope as a global descriptor capable of capturing the 'gist' of a scene and demonstrate that it can be used to distinguish among eight different categories of natural scenes has sparked tremendous interest in the computer vision research community and can be considered the seminal reference at the beginning of the most recent wave of work on the topic of context modeling. Other essential papers include [45, 53, 80, 81, 95, 98, 110–116].

Many project-related resources are also available online, including:

- MATLAB code, datasets, and examples of results for the "spatial envelope" scene representation [79]: http://people.csail.mit.edu/torralba/code/spatialenvelope/. Images for each of the eight scene categories can also be downloaded from http://cvcl.mit.edu/database.htm.

[2] http://web.mit.edu/torralba/www/

[3] http://cvcl.mit.edu/Aude.htm

- Datasets and examples of results for the "Place and scene recognition from video" project [114]: http://www.cs.ubc.ca/~murphyk/Vision/placeRecognition.html.
- Datasets, tools, and examples of results for contextual priming for object detection [117]: http://web.mit.edu/torralba/www/carsAndFacesInContext.html.
- MATLAB code, datasets, tools, and examples of results for contextual guidance of eye movements and attention in real-world scenes [116]: http://people.csail.mit.edu/torralba/GlobalFeaturesAndAttention/.

Efros, Hoiem et al.

The research by Efros,[4] Hoiem (currently at University of Illinois at Urbana-Champaign)[5] and colleagues at Carnegie Mellon University has led to some of the most influential practical applications of the latest computer vision techniques to the problem of modeling and reconstructing 3D scenes and using contextual information to improve the performance of object detection and recognition solutions.

Essential reading include [55–60]. Many project-related resources are also available, including:

- MATLAB code, datasets, and examples of results for the "Automatic Photo Pop-up" project [56]: http://www.cs.uiuc.edu/homes/dhoiem/projects/popup/index.html
- MATLAB code, datasets, and examples of results for the "Surface Context" project [57, 58]: http://www.cs.uiuc.edu/homes/dhoiem/projects/context/index.html
- Dataset for the "Putting Objects in Perspective" project [59]: http://www.cs.uiuc.edu/homes/dhoiem/projects/pop/index.html
- Code and ground-truth data for the "Recovering the Spatial Layout of Cluttered Rooms" project [49]: https://netfiles.uiuc.edu/vhedau2/www/Research/research_spatialLayout.html

The Make3D team

The Make 3D project is another approach to learn depth and infer 3D model from a single image. It was created by Saxena, Ng and their colleagues from Stanford 3D reconstruction group [97]. Code and range image data for Make3D are available at http://make3d.cs.cornell.edu/code.html.

Heitz and Koller

Heitz and Koller (Stanford University) are the proponents of the "Things and stuff" (TAS) context model discussed earlier (Sections 3.2.2 and 4.3.2). Code and image data associated with their work can be found at: http://ai.stanford.edu/ gaheitz/Research/TAS/.

[4] http://www.cs.cmu.edu/~efros/

[5] http://www.cs.uiuc.edu/homes/dhoiem/

Grauman et al.

The work by Grauman[6] and colleagues at University of Texas at Austin covers a broad range of topics, from "learning and recognizing visual object categories" to "scalable methods for content-based retrieval and visual search".

Project-related resources include:

- Datasets and examples of results for the "Reading Between The Lines: Object Localization Using Implicit Cues from Image Tags" project [63]: http://userweb.cs.utexas.edu/~sjhwang/tags.html
- MATLAB code, supplementary materials, and examples of results for the "Object-Graphs for Context-Aware Category Discovery" project [71]: http://userweb.cs.utexas.edu/~grauman/research/projects/objectgraph/objectgraph.html

Belongie, Rabinovich, Galleguillos, et al.

The work by Belongie, Rabinovich, Galleguillos and colleagues at the Computer Vision Laboratory in the Computer Science and Engineering Department at U.C. San Diego has produced significant impact on the state of the art in the topic of "Context Based Object Categorization". Recommended reading include [39–41, 92, 93].

Leonardis et al.

Leonardis[7] and colleagues at the Visual Cognitive Systems Laboratory at the University of Ljubljana have been working in the field of context-aware object detection. For demonstration videos associated with the projects reported in [87], visit: http://vicos.fri.uni-lj.si/roli/research/context-aware-object-detection/. Three urban image datasets (Ljubljana, Graz, and Darmstadt) can be downloaded from: http://vicos.fri.uni-lj.si/downloads/.

Datasets

Experimental research on computational approaches for contextual modeling, object, and scene recognition, should be extensively tested using representative images. In recent years, several image collections have been made publicly available to the research community, which brings many advantages such as time savings (capturing, selecting, organizing, and annotating images is a very time consuming process) and the ability to compare (i.e., benchmark) a new algorithm or framework against previous approaches in the literature. In addition to proprietary image and video collections, several recent experiments in this field have used one of the following publicly available datasets:

- PASCAL Visual Object Classes (VOC) dataset [31] http://pascallin.ecs.soton.ac.uk/challenges/VOC/

[6] http://userweb.cs.utexas.edu/~grauman/

[7] http://vicos.fri.uni-lj.si/alesl/

The PASCAL VOC dataset consists of consumer photographs collected from Flickr and associated ground truth annotation (including coordinates of the rectangular areas delimiting an object of interest). The images are used in the context of two principal challenges: object classification and object detection. New datasets have been released each year since 2006. The images are organized into 20 classes as follows:

- Person: person
- Animal: bird, cat, cow, dog, horse, sheep
- Vehicle: airplane, bicycle, boat, bus, car, motorbike, train
- Indoor: bottle, chair, dining table, potted plant, sofa, TV/monitor

The PASCAL VOC dataset has been used by Divvala and colleagues [27] and Heitz and Koller [51], among others.

In spite of its great popularity and usefulness, the PASCAL dataset has been deemed not suitable for experiments with context-based object recognition algorithms by Choi et al. [18], because most images contain very few instances of a single object category (more than 50% of the images contain only a single object class) and also because objects' bounding boxes occupy a large portion (typically 20%) of the image.

- LabelMe dataset [96] http://labelme.csail.mit.edu/

  The LabelMe dataset consists of an ever-growing collection of 180,000+ images and associated annotations, contributed by its users in a collaborative way. The images—and associated MATLAB code to process, query, and annotate them—are publicly available and cover a wide range of topics and scenarios. One of the main criticisms of the LabelMe dataset refers to the fact that the dataset is incompletely labeled, since volunteer annotators are free to choose which objects to annotate, and which to omit, leading to difficulties in establishing precision and recall for detection and classification tasks [31]. Consequently, researchers interested in using LabelMe for their experimental evaluations typically adopt selected subsets of the database to use for training and testing, and ensure that these subsets are completely annotated. Subsets of the LabelMe dataset have been used by Oliva and Torralba [81], among others.
- SUN 09 dataset [18] http://web.mit.edu/~myungjin/www/HContext.html

  A few months ago, a new dataset was proposed and made available to the research community: the SUN 09 dataset, which contains 12,000 annotated images covering a large number of scene categories (indoor and outdoors) with more than 200 object categories and 152,000 annotated object instances. SUN 09 contains images collected from multiple sources (Google, Flickr, Altavista, LabelMe) and does not include images of objects on white backgrounds or close-ups, i.e., images in which there is no significant context information. It has been annotated using LabelMe [96] by a single annotator and verified for consistency [18].

  In the SUN 09 dataset, the average object size is 5% of the image size, and a typical image contains seven different object categories. In their evaluation, Choi et al. demonstrate that SUN 09 contains richer contextual information when compared to PASCAL VOC 2007, using the same 20 categories. They also demonstrate that the contextual information learned from SUN 09 significantly improves the accuracy of object recognition tasks, and can even be used to identify out-of-context (e.g., due to wrong pose, scale, or co-occurrence) scenes [18].

- SUN dataset [123] http://groups.csail.mit.edu/vision/SUN/
  The Scene UNderstanding (SUN) dataset was introduced earlier this year and is targeted at research in scene classification, which has been customarily tested on a fairly small (usually, 15 or less) number of semantic categories. The SUN dataset contains 899 categories and 130,519 images. The number of images varies across categories, but there are at least 100 images per category. Out of the 899 categories, in [123], the authors use 397 well-sampled categories to evaluate numerous state-of-the-art algorithms for scene recognition and establish new bounds of performance. All images, associated code, as well as training and testing partitions are available for download at the URL indicated above.
- NUS-WIDE [21] http://lms.comp.nus.edu.sg/research/NUS-WIDE.htm
  The NUS-WIDE is a publicly accessible image dataset created by the Lab for Media Search at National University of Singapore (NUS). The dataset includes:

  1. 269,648 images and the associated tags from Flickr, with a total number of 5,018 unique tags;
  2. six types of low-level features extracted from these images, including 64-D color histogram, 144-D color correlogram, 73-D edge direction histogram, 128-D wavelet texture, 225-D block-wise color moments and 500-D bag of words based on SIFT descriptions; and
  3. ground-truth for 81 concepts that can be used for evaluation.

Additionally, there are several web pages with lists of links to useful datasets for computer vision, including:

- http://userweb.cs.utexas.edu/~grauman/courses/spring2008/datasets.htm—maintained by Prof. Kristen Grauman (University of Texas at Austin)
- http://www.cs.ubc.ca/~murphyk/Vision/objectRecognitionDatabases.html—maintained by Prof. Kevin Murphy (University of British Columbia)
- http://www.cs.cmu.edu/~efros/courses/LBMV07/databases.htm—maintained by Prof. Alexei Efros (Carnegie Mellon University)

References

1. Alvarez G, Oliva A (2008) The representation of simple ensemble visual features outside the focus of attention. Psychol Sci 19(4):392–398
2. Amores J, Radeva P (2005) Registration and retrieval of highly elastic bodies using contextual information. Pattern Recogn Lett 26(11):1720–1731
3. Amores J, Sebe N, Radeva P (2007) Context-based object-class recognition and retrieval by generalized correlograms. IEEE Trans Pattern Anal Mach Intell 29(10):1818–1833
4. Ariely D (2001) Seeing sets: representation by statistical properties. Psychol Sci 12(2):157–162
5. Auckland M (2007) Non-target objects can influence perceptual processes during object recognition. Psychon Bull Rev 14:332–337
6. Bar M (2004) Visual objects in context. Nat Rev Neurosci 5(8):617–629
7. Bar M, Aminoff E (2003) Cortical analysis of visual context. Neuron 38(2):347–358
8. Bar M, Ullman S (1996) Spatial context in recognition. Perception 25(3):343–352

9. Barenholtz E (2009) Quantifying the role of context in visual object recognition [abstract]. J Vis 9(8):800, 800a
10. Belongie S, Malik J, Puzicha J (2002) Shape matching and object recognition using shape contexts. IEEE Trans Pattern Anal Mach Intell 24:509–522
11. Biederman I, Mezzanote R, Rabinovitz J (1982) Scene perception: detecting and judging objects undergoing relational violations. Cogn Psychol 14:143–147
12. Biederman I, Rabinowitz JC, Glass AL, Stacy EW (1974) On the information extracted from a glance at a scene. J Exp Psychol 103(3):597–600
13. Boyce JS, Pollatsek A, Rayner K (1998) Effect of background information on object identification. J Exp Psychol Hum Percept Perform 15(3):556–566
14. Brockmole JR, Castelhano MS, Henderson JM (2006) Contextual cueing in naturalistic scenes: global and local contexts. J Exp Psychol Learn Mem Cogn 32(4):699–706
15. Brockmole JR, Hambrick DZ, Windisch DJ, Henderson JM (2008) The role of meaning in contextual cueing: evidence from chess expertise. Q J Exp Psychol (Colchester) 61(12):1886–1896
16. Cao L, Luo J, Kautz H, Huang T (2009) Image annotation within the context of personal photo collections using hierarchical event and scene models. IEEE Trans Multimedia 11(2): 208–219
17. Carbonetto P, de Freitas N, Barnard K (2004) A statistical model for general contextual object recognition. In: European conference on computer vision (ECCV), pp 350–362
18. Choi MJ, Lim J, Torralba A, Willsky A (2010) Exploiting hierarchical context on a large database of object categories. In: Computer vision and pattern recognition (CVPR), 2010 IEEE conference on, pp 129–136
19. Chong SC, Treisman A (2005) Attentional spread in the statistical processing of visual displays. Percept Psychophys 67(1):1–13
20. Chong SC, Treisman A (2005) Statistical processing: computing the average size in perceptual groups. Vis Res 45(7):891–900
21. Chua T-S, Tang J, Hong R, Li H, Luo Z, Zheng Y-T (2009) Nus-wide: a real-world web image database from National University of Singapore. In: Proc. of ACM conf. on image and video retrieval (CIVR'09). Santorini, Greece
22. Chun M, Jiang Y (1999) Top-down attentional guidance based on implicit learning of visual covariation. Psychol Sci 10:360–365
23. Chun M, Jiang Y (2003) Implicit, long-term spatial contextual memory. Percept Psychophys 65:72–80
24. Chun MM, Jiang Y (1998) Contextual cueing: implicit learning and memory of visual context guides spatial attention. Cogn Psychol 36(1):28–71
25. Cox D, Meyers E, Sinha P (2004) Contextually evoked object-specific responses in human visual cortex. Science 304:115–117
26. Davenport JL, Potter MC (2004) Scene consistency in object and background perception. Psychol Sci 15(8):559–564
27. Divvala S, Hoiem D, Hays J, Efros A, Hebert M (2009) An empirical study of context in object detection. In: Computer vision and pattern recognition recognition, CVPR 2009. IEEE conference on, pp 1271–1278
28. Endo N, Takeda Y (2005) Use of spatial context is restricted by relative position in implicit learning. Psychon Bull Rev 12(5):880–885
29. Epstein R, Harris A, Stanley D, Kanwisher N (1999) The parahippocampal place area: recognition, navigation, or encoding? Neuron 23(1):115–125
30. Epstein R, Kanwisher N (1998) A cortical representation of the local visual environment. Nature 392:598–601
31. Everingham M, Van Gool L, Williams CKI, Winn J, Zisserman A (2010) The Pascal Visual Object Classes (VOC) challenge. Int J Comput Vis 88(2):303–338
32. Fei-Fei L, Iyer A, Koch C, Perona P (2007) What do we perceive in a glance of a real-world scene? J Vis 7(1):1–29
33. Felzenszwalb P, Girshick R, McAllester D (2010) Cascade object detection with deformable part models. In: Computer vision and pattern recognition (CVPR), 2010 IEEE conference on, pp 2241–2248
34. Felzenszwalb PF, Huttenlocher DP (2004) Efficient graph-based image segmentation. Int J Comput Vis 59:167–181
35. Ferrari V, Jurie F, Schmid C (2010) From images to shape models for object detection. Int J Comput Vis 87(3):284–303

36. Fink M, Perona P (2003) Mutual boosting for contextual inference. In: Thrun S, Saul L, Schökopf B (eds) Advances in neural information processing systems (NIPS). MIT Press, Cambridge, MA
37. Fischler MA, Strat TM (1989) Recognizing objects in a natural environment: a contextual vision system (cvs). In: Proceedings of a workshop on image understanding workshop. San Francisco, CA, USA. Morgan Kaufmann, pp 774–796
38. Forsyth D, Malik J, Fleck M, Greenspan H, Leung T, Belongie S, Carson C, Bregler C (1996) Finding pictures of objects in large collections of images. Technical report, UC Berkeley, Berkeley, CA, USA
39. Galleguillos C, Belongie S (2010) Context based object categorization: a critical survey. Comput Vis Image Underst (CVIU) 114:712–722
40. Galleguillos C, McFee B, Belongie S, Lanckriet GRG (2010) Multi-class object localization by combining local contextual interactions. In: IEEE conference in computer vision and patter recognition (CVPR)
41. Galleguillos C, Rabinovich A, Belongie S (2008) Object categorization using co-occurrence, location and appearance. In: Proc. IEEE conf. computer vision and pattern recognition (CVPR), pp 1–8
42. Goujon A, Didierjean A, Marmèche E (2007) Contextual cueing based on specific and categorical properties of the environment. Vis Cogn 15:257–275
43. Goujon A, Didierjean A, Marmèche E (2009) Semantic contextual cuing and visual attention. J Exp Psychol Hum Percept Perform 35(1):50–71
44. Graef PD, Troy AD, D'Ydewalle G (1992) Local and global contextual constraints on the identification of objects in scenes. Can J Psychol 46(3):489–508
45. Greene M, Oliva A (2009) Recognition of natural scenes from global properties: seeing the forest without representing the trees. Cogn Psychol 58:137–176
46. Gronau N, Neta M, Bar M (2008) Integrated contextual representation for objects' identities and their locations. J Cogn Neurosci 20(3):371–388
47. Gupta A, Efros AA, Hebert M (2010) Blocks world revisited: image understanding using qualitative geometry and mechanics, in ECCV
48. Hays J, Efros A (2008) IM2GPS: estimating geographic information from a single image. In: Computer vision and pattern recognition, CVPR 2008. IEEE conference on, pp 1–8
49. Hedau V, Hoiem D, Forsyth D (2009) Recovering the spatial layout of cluttered rooms. In: Computer vision, 2009 IEEE 12th international conference on, pp 1849–1856
50. Hedau V, Hoiem D, Forsyth D (2010) Thinking inside the box: using appearance models and context based on room geometry. In: Daniilidis K, Maragos P, Paragios N (eds) Computer vision – ECCV 2010. Springer Berlin, Heidelberg, pp 224–237
51. Heitz G, Koller D (2008) Learning spatial context: using stuff to find things. In: ECCV '08: Proceedings of the 10th European conference on computer vision. Springer, Berlin, pp 30–43
52. Henderson JM, Hollingworth A (1999) High-level scene perception. Annu Rev Psychol 50: 243–271
53. Hidalgo-Sotelo B, Oliva A, Torralba A (2005) Human learning of contextual priors for object search: where does the time go? In: Proceedings of the IEEE Computer Society conference on computer vision and pattern recognition, vol 3, pp 86–93
54. Hock H (1974) Contextual relations: the influence of familiarity, physical plausibility, and belongingness. Percept Psychophys 16:4–8
55. Hoiem D, Efros A, Hebert M (2008) Closing the loop in scene interpretation. In: Computer vision and pattern recognition, CVPR 2008. IEEE conference on, pp 1–8
56. Hoiem D, Efros AA, Hebert M (2005) Automatic photo pop-up. ACM Trans Graph (SIGGRAPH 2005) 24(3):577–584
57. Hoiem D, Efros AA, Hebert M (2005) Geometric context from a single image. In: ICCV '05: Proceedings of the tenth IEEE international conference on computer vision (ICCV'05), vol 1. IEEE Computer Society, Washington, pp 654–661
58. Hoiem D, Efros AA, Hebert M (2007) Recovering surface layout from an image. Int J Comput Vis 75(1):151–172
59. Hoiem D, Efros AA, Hebert M (2008) Putting objects in perspective. Int J Comput Vis 80(1): 3–15
60. Hoiem D, Stein A, Efros A, Hebert M (2007) Recovering occlusion boundaries from a single image. In: Computer vision, ICCV 2007. IEEE 11th international conference on, pp 1–8

61. Hollingworth A, Henderson JM (1998) Does consistent scene context facilitate object perception? J Exp Psychol Gen 127(4):398–415
62. Huang J, Kumar S, Mitra M, Zhu W-J, Zabih R (1997) Image indexing using color correlograms. In: Computer vision and pattern recognition. Proceedings, 1997 IEEE computer society conference on, pp 762–768
63. Hwang SJ, Grauman K (2010) Reading between the lines: object localization using implicit cues from image tags. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR). San Francisco, CA
64. Itti L, Koch C (2001) Computational modelling of visual attention. Nat Rev Neurosci 2(3):194–203
65. Joshi D, Luo J (2008) Inferring generic activities and events from image content and bags of geotags. In: CIVR '08: Proceedings of the 2008 international conference on content-based image and video retrieval. ACM, New York, pp 37–46
66. Kennedy LS, Naaman M (2008) Generating diverse and representative image search results for landmarks. In: WWW '08: Proceeding of the 17th international conference on World Wide Web. ACM, New York, pp 297–306
67. Koch C, Ullman S (1985) Shifts in selective visual attention: towards the underlying neural circuitry. Hum Neurobiol 4(4):219–227
68. Kumar M, Torr P, Zisserman A (2005) OBJ CUT. In: Computer vision and pattern recognition, CVPR 2005. IEEE computer society conference on, vol 1, pp 18–25
69. Kumar S, Hebert M (2005) A hierarchical field framework for unified context-based classification. In: ICCV '05: Proceedings of the tenth IEEE international conference on computer vision. IEEE Computer Society, Washington, pp 1284–1291
70. Kunar M (2007) Does contextual cueing guide the deployment of attention? J Exp Psychol Hum Percept Perform 33:816–828
71. Lee YJ, Grauman K (2010) Object-graphs for context-aware category discovery. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR). San Francisco, CA
72. Leordeanu M, Hebert M, Sukthankar R (2007) Beyond local appearance: category recognition from pairwise interactions of simple features. In: Computer vision and pattern recognition, CVPR '07. IEEE conference on, pp 1–8
73. Luo J, Yu J, Joshi D, Hao W (2008) Event recognition: viewing the world with a third eye. In: MM '08: Proceeding of the 16th ACM international conference on multimedia. ACM, New York, pp 1071–1080
74. Marr D (1982) Vision. W. H. Freeman, San Francisco
75. Modestino J, Zhang J (1992) A Markov random field model-based approach to image interpretation. IEEE Transactions on Pattern Anal Mach Intell 14:606–615
76. Navon D (1977) Forest before trees: the precedence of global features in visual perception. Cogn Psychol 9:353–383
77. O'Hare N, Lee H, Cooray S, Gurrin C, Jones G, Malobabic J, O'Connor N, Smeaton AF, Uscilowski B (2006) Mediassist: using content-based analysis and context to manage personal photo collections. In: 5th int. conf. on image and video retrieval. Tempe, AZ, pp 529–532
78. O'Hare N, Smeaton A (2009) Context-aware person identification in personal photo collections. IEEE Trans Multimedia 11(2):220–228
79. Oliva A, Torralba A (2001) Modeling the shape of the scene: a holistic representation of the spatial envelope. Int J Comput Vis 42(3):145–175
80. Oliva A, Torralba A (2006) Building the gist of a scene: the role of global image features in recognition. Prog Brain Res 155:23–36
81. Oliva A, Torralba A (2007) The role of context in object recognition. Trends Cogn Sci 11(12):520–527
82. Opelt A, Pinz A, Zisserman A (2006) A boundary-fragment-model for object detection. In: Leonardis A, Bischof H, Pinz A (eds) Computer vision – ECCV 2006. Springer Berlin, Heidelberg, pp 575–588
83. Opelt A, Pinz A, Zisserman A (2006) Incremental learning of object detectors using a visual shape alphabet. In: Computer vision and pattern recognition, 2006 IEEE computer society conference on, pp 3–10
84. Opelt A, Pinz A, Zisserman A (2008) Learning an alphabet of shape and appearance for multiclass object detection. Int J Comput Vis 80(1):16–44

85. Palmer S (1975) The effects of contextual scenes on the identification of objects. Mem Cogn 3:519–526
86. Peissig J, Tarr M (2007) Visual object recognition: do we know more now than we did 20 years ago? Annu Rev Psychol 50:75–96
87. Perko R, Leornardis A (2010) A framework for visual-context-aware object detection in still images. Comput Vis Image Underst (CVIU) 114:700–711
88. Posner MI (1980) Orienting of attention. Q J Exp Psychol 32(1):3–25
89. Potter MC (1976) Short-term conceptual memory for pictures. J Exp Psychol Hum Learn 2(5):509–522
90. Potter MC, Faulconer BA (1975) Time to understand pictures and words. Nature 253(5491): 437–438
91. Quattoni A, Torralba A (2009) Recognizing indoor scenes. In: Computer vision and pattern recognition, CVPR 2009. IEEE conference on, pp 413–420
92. Rabinovich A, Belongie S (2009) Scenes vs. objects: a comparative study of two approaches to context based recognition. In: Computer vision and pattern recognition workshops. CVPR workshops 2009. IEEE computer society conference on, pp 92–99
93. Rabinovich A, Vedaldi A, Galleguillos C, Wiewiora E, Belongie S (2007) Objects in context. In: Computer vision. ICCV 2007. IEEE 11th international conference on, pp 1–8
94. Rieger JW, Köchy N, Schalk F, Grüschow M, Heinze H-J (2008) Speed limits: orientation and semantic context interactions constrain natural scene discrimination dynamics. J Exp Psychol Hum Percept Perform 34(1):56–76
95. Russell B, Torralba A, Liu C, Fergus R, Freeman W (2007) Object recognition by scene alignment. In: Platt JC, Koller D, Singer Y, Roweis S (eds) Advances in neural information processing systems (NIPS). MIT Press, Cambridge, MA, pp 1241–1248
96. Russell BC, Torralba A, Murphy KP, Freeman WT (2008) Labelme: a database and web-based tool for image annotation. Int J Comput Vis 77:157–173
97. Saxena A, Sun M, Ng AY (2009) Make3d: learning 3d scene structure from a single still image. IEEE Trans Pattern Anal Mach Intell 31(5):824–840
98. Schyns P, Oliva A (1994) From blobs to boundary edges: evidence for time- and spatial-scale-dependent scene recognition. Psychol Sci 5:195–200
99. Selfridge OG (1955) Pattern recognition and modern computers. In: Proceedings of the western joint computer conference. IEEE, New York
100. Siagian C, Itti L (2007) Rapid biologically-inspired scene classification using features shared with visual attention. IEEE Trans Pattern Anal Mach Intell 29(2):300–312
101. Singhal A, Luo J, Zhu W (2003) Probabilistic spatial context models for scene content understanding. In: Computer vision and pattern recognition. Proceedings, 2003 IEEE computer society conference on, vol 1, pp I-235–I-241
102. Smeulders A, Worring M, Santini S, Gupta A, Jain R (2000) Content-based image retrieval at the end of the early years. IEEE Trans Pattern Anal Mach Intell 22(12):1349–1380
103. Strat T (1992) Natural object recognition. Springer-Verlag New York, Inc., New York, NY, USA
104. Strat T (1993) Employing contextual information in computer vision. In: Proceedings of ARPA image understanding workshop, pp 217–229
105. Strat T, Fischler M (1989) Context-based vision: recognition of natural scenes. In: Twenty-third asilomar conference on signals, systems and computers, pp 532–536
106. Strat T, Fischler M (1990) A context-based recognition system for natural scenes and complex domains. In: DARPA image understanding workshop, pp 456–472
107. Strat T, Fischler M (1991) Context-based vision: recognizing objects using information from both 2-d and 3-d imagery. IEEE Trans Pattern Anal Mach Intell 13(10):1050–1065
108. Strat T, Fua P, Connolly C (1997) Context-based vision. In: Radius: image understanding for imagery intelligence, pp 373–388
109. Thorpe S, Fize D, Marlot C (1996) Speed of processing in the human visual system. Nature 381(6582):520–522
110. Torralba A (2003) Contextual priming for object detection. Int J Comput Vis 53:169–191
111. Torralba A (2003) Modeling global scene factors in attention. J Opt Soc Am A 20(7):1407–1418
112. Torralba A, Murphy KP, Freeman W (2004) Contextual models for object detection using boosted random fields. In Saul LK, Weiss Y, Bottou L (eds) Advances in neural information processing systems (NIPS). MIT Press, Cambridge, MA, pp 1401–1408
113. Torralba A, Murphy KP, Freeman WT (2010) Using the forest to see the trees: exploiting context for visual object detection and localization. Commun ACM 53(3):107–114

114. Torralba A, Murphy KP, Freeman WT, Rubin MA (2003) Context-based vision system for place and object recognition. In: Proc ninth IEEE int computer vision conf, pp 273–280
115. Torralba A, Oliva A (2003) Statistics of natural image categories. Network 14(3):391–412
116. Torralba A, Oliva A, Castelhano MS, Henderson JM (2006) Contextual guidance of eye movements and attention in real-world scenes: the role of global features in object search. Psychol Rev 113(4):766–786
117. Torralba A, Sinha P (2001) Statistical context priming for object detection. In: Computer vision. ICCV 2001. Proceedings eighth IEEE international conference on, pp 763–770
118. Treisman AM, Gelade G (1980) A feature-integration theory of attention. Cogn Psychol 12(1):97–136
119. Viola P, Jones MJ (2004) Robust real-time face detection. Int J Comput Vision 57(2):137–154
120. Zheng W-S, Gong S, Xiang T (2009) Quantifying contextual information for object detection. In: Computer vision, 2009 IEEE 12th international conference on, pp 932–939
121. Wang X, Doretto G, Sebastian T, Rittscher J, Tu PH (2007) Shape and appearance context modeling. In: IEEE 11th international conference on computer vision (ICCV) 2007, 14–21 Oct 2007. Rio de Janeiro, Brazil, pp 1–8
122. Wolf L, Bileschi S (2006) A critical view of context. Int J Comput Vision 69(2):251–261
123. Xiao J, Hays J, Ehinger K, Oliva A, Torralba A (2010) SUN database: large-scale scene recognition from abbey to zoo. In: Computer vision and pattern recognition (CVPR), 2010 IEEE conference on, pp 3485–3492
124. Yakimovsky Y, Feldman JA (1973) A semantics-based decision theory region analyzer. In: IJCAI'73: Proceedings of the 3rd international joint conference on artificial intelligence. Morgan Kaufmann, San Francisco, pp 580–588
125. Yang Y, Hallman S, Ramanan D, Fowlkes C (2010) Layered object detection for multi-class segmentation. In: Computer vision and pattern recognition (CVPR), 2010 IEEE conference on, pp 3113–3120
126. Yang YH, Wu PT, Lee CW, Lin KH, Hsu WH, Chen HH (2008) Contextseer: context search and recommendation at query time for shared consumer photos. In: MM '08: Proceeding of the 16th ACM international conference on multimedia. ACM, New York, pp 199–208
127. Yantis S (1993) Stimulus-driven attentional capture and attentional control settings. J Exp Psychol Hum Percept Perform 19(3):676–681
128. Yantis S, Jonides J (1990) Abrupt visual onsets and selective attention: voluntary versus automatic allocation. J Exp Psychol Hum Percept Perform 16(1):121–134

Oge Marques is an Associate Professor and Associate Chairman in the Department of Computer Science and Engineering at Florida Atlantic University in Boca Raton, Florida. He received his Ph.D. in Computer Engineering from Florida Atlantic University in 2001, his Master's in Electronics Engineering from Philips International Institute (Eindhoven, Netherlands) in 1989 and his Bachelor's

Degree in Electrical Engineering from UTFPR (Curitiba, Brazil) in 1987. His research interests include: image processing, analysis, annotation, search, and retrieval; human and computer vision; and video processing and analysis. He has more than 20 years of teaching and research experience in the fields of image processing and computer vision, in different countries and capacities. He is the (co-) author of four books in these topics, including the forthcoming textbook "Practical Image and Video Processing Using MATLAB" (Wiley, 2011). He has also published several book chapters and more than 50 refereed journal and conference papers in these fields. He serves as a reviewer and Editorial Board member for several leading journals in computer science and engineering. He is a senior member of the ACM, senior member of the IEEE, and a member of the IEEE Computer Society, IEEE Education Society, and the honor societies of Tau Beta Pi, Sigma Xi, Phi Kappa Phi and Upsilon Pi Epsilon.

Elan Barenholtz is an Assistant Professor in the Department of Psychology at Florida Atlantic University in Boca Raton, Florida. He received his Ph.D. in Cognitive Psychology, from Rutgers University/New Brunswick in 2004. He then worked as a postdoctoral research fellow at Brown University in Cognitive Science before joining the faculty of FAU in 2008. His research interests include visual shape perception, scene and object recognition and multimodal integration. He is currently focusing his research on contextual facilitation of object recognition under a grant received from the National Science Foundation entitled: "Identifying Objects Within Scenes: Combining Context and Features in Visual Object Recognition". He is a member of the Vision Sciences Society, the Psychonomic Society and The American Psychological Association.

Vincent Charvillat received the Eng. degree in Computer Science and Applied Mathematics from ENSEEIHT, Toulouse France and the M.Sc. in Computer Science from the National Polytechnic Institute of Toulouse, both in 1994. He received the Ph.D. degree in Computer Science from the National Polytechnic Institute of Toulouse in 1997. He joined the Computer Science and Applied Mathematics department of ENSEEIHT in 1998 as an assistant professor. He obtained the habilitation degree in Computer Science in 2008 and is currently a professor at the University of Toulouse, IRIT research lab, ENSEEIHT Eng. School. Vincent Charvillat is the head of VORTEX research team at ENSEEIHT (Visual Objects: from Reality To EXpression). His main research interests are visual processing and multimedia applications. Current topics of research include visual object extraction (object tracking, detection and coding), compositing (augmented reality and hypermedia), interactive delivery (multimedia adaptation, mobile applications).