Bogdan Ionescu
Wilma A. Bainbridge
Naila Murray *Editors*

Human Perception of Visual Information

Psychological and Computational Perspectives

Springer

Human Perception of Visual Information

Bogdan Ionescu • Wilma A. Bainbridge
Naila Murray
Editors

Human Perception of Visual Information

Psychological and Computational Perspectives

Springer

Editors
Bogdan Ionescu
Politehnica University of Bucharest
Bucharest, Romania

Wilma A. Bainbridge
University of Chicago
Chicago, IL, USA

Naila Murray
NAVER Labs Europe
Meylan, France

ISBN 978-3-030-81464-9 ISBN 978-3-030-81465-6 (eBook)
https://doi.org/10.1007/978-3-030-81465-6

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Preface

There is one thing the photograph must contain, the humanity of the moment.

—Robert Frank

Computational models of objective visual properties such as semantic content and geometric relationships have made significant breakthroughs using the latest achievements in machine learning and large-scale data collection. There has also been limited but important work exploiting these breakthroughs to improve computational modelling of subjective visual properties such as interestingness, affective values and emotions, aesthetic values, memorability, novelty, complexity, visual composition and stylistic attributes, and creativity. Researchers that apply machine learning to model these subjective properties are often motivated by the wide range of potential applications of such models, including for content retrieval and search, storytelling, targeted advertising, education and learning, and content filtering. The performance of such machine learning-based models leaves significant room for improvement and indicates a need for fundamental breakthroughs in our approach to understanding such highly complex phenomena.

Largely in parallel to these efforts in the machine learning community, recent years have witnessed important advancements in our understanding of the psychological underpinnings of these same subjective properties of visual stimuli. Early focuses in the vision sciences were on the processing of simple visual features like orientations, eccentricities, and edges. However, utilizing new neuroimaging techniques such as functional magnetic resonance imaging, breakthroughs through the 1990s and 2000s uncovered specialized processing in the brain for high-level visual information, such as image categories (e.g., faces, scenes, tools, objects) and more complex image properties (e.g., real-world object size, emotions, aesthetics). Recent work in the last decade has leveraged machine learning techniques to allow researchers to probe the specific content of visual representations in the brain. In parallel, the widespread advent of the Internet has allowed for large-scale crowd-sourced experiments, allowing psychologists to go beyond small samples with limited, controlled stimulus sets to study images at a large scale. With the combination of these advancements, psychology is now able to take a fresh look at

age-old questions like what we find interesting, what we find beautiful, what drives our emotions, how we perceive spaces, or what we remember.

The field of machine learning, and Artificial Intelligence more broadly, enjoys a long tradition of seeking inspiration from investigations into the psychology and neuroscience of human and non-human intelligence. For example, deep learning neural networks in Computer Vision were originally inspired by the architecture of the human visual system, with its many layers of neurons thought to apply filters at each stage. Psychology and neuroscience also rely heavily on developments from Artificial Intelligence, both for parsing down the Big Data collected from the brain and behavior, as well as for understanding the underlying mechanisms. For example, now, object classification deep neural networks such as VGG-16 are frequently used as stand-ins for the human visual system to predict behavior or even activity in the brain. Given the progress made in machine learning and psychology towards more successfully modelling subjective visual properties, we believe that the time is ripe to explore how these advances can be mutually enriching and lead to further progress.

To that end, this book showcases complementary perspectives from psychology and machine learning on high-level perception of images and videos. It is an interdisciplinary volume that brings together experts from psychology and machine learning in an attempt to bring these two, at a first glance, different fields, into conversation, while at the same time providing an overview of the state of the art in both fields. The book contains 10 chapters arranged in 5 pairs, with each pair describing state-of-the-art psychological and computational approaches to describing and modelling a specific subjective perceptual phenomenon.

In Chap. 1, Lauer and Võ review recent studies that use diverse methodologies like psychophysics, eye tracking, and neurophysiology to help better capture human efficiency in real-world scene and object perception. The chapter focuses in particular on which contextual information humans take advantage of most and when. Further, they explore how these findings could be useful in advancing computer vision and how computer vision could mutually further understanding of human visual perception. In Chap. 2, Constantin et al. consider the related phenomenon of interestingness prediction from a computational point of view and present an overview of traditional fusion mechanisms, such as statistical fusion, weighted approaches, boosting, random forests, and randomized trees. They also include an investigation of a novel, deep learning-based system fusion method for enhancing performance of interestingness prediction systems.

In Chap. 3, Bradley et al. review recent research related to photographic images that depict affectively engaging events, with the goal of assessing the extent to which specific pictures reliably engage emotional reactions across individuals. In particular, they provide preliminary analyses that encourage future investigations aimed at constructing normative biological image databases that, in addition to evaluative reports, provide estimates of emotional reactions in the body and brain for use in studies of emotion and emotional dysfunction. On the computational side, in Chap. 4, Zhao et al. introduce image emotion analysis from a computational perspective with a focus on summarizing recent advances. They revisit key computational

problems with emotion analysis and present in detail aspects such as emotion feature extraction, supervised classifier learning, and domain adaptation. Their discussion concludes with the presentation of the relevant datasets for evaluation and the identification of open research directions.

In Chap. 5, Chamberlain sets out the history of empirical aesthetics in cognitive science and the state of the research field at present. The chapter outlines recent work on inter-observer agreement in aesthetic preference before presenting empirical work that argues the importance of objective (characteristics of stimuli) and subjective (characteristics of context) factors in shaping aesthetic preference. Valenzise et al. explore machine learning approaches to modelling computational image aesthetics, in Chap. 6. They overview the several interpretations that aesthetics have received over time and introduce a taxonomy of aesthetics. They discuss computational advances in aesthetics prediction, from early methods to deep neural networks, and overview the most popular image datasets. Open challenges are identified and discussed, including dealing with the intrinsic subjectivity of aesthetic scores and providing explainable aesthetic predictions.

Bainbridge, in Chap. 7, draws from neuroimaging and other research to describe our current state-of-the-art understanding of memorability of visual information. Such research has revealed that the brain is sensitive to memorability both rapidly and automatically during late perception. These strong consistencies in memory across people may reflect the broad organizational principles of our sensory environment and may reveal how the brain prioritizes information before encoding items into memory. In Chap. 8, Bylinskii et al. examine the notion of memorability with a computational lens, detailing the state-of-the-art algorithms that accurately predict image memorability relative to human behavioral data, using image features at different scales from raw pixels to semantic labels. Beyond prediction, they show how recent Artificial Intelligence approaches can be used to create and modify visual memorability, and preview the computational applications that memorability can power, from filtering visual streams to enhancing augmented reality interfaces.

In Chap. 9, Akcelik et al. review recent research that aims to quantify visual characteristics and design qualities of built environments, in order to relate more abstract aspects of an urban space to quantifiable design features. Uncovering these relationships may provide the opportunity to establish a causal relationship between design features and psychological feelings such as walkability, preference, visual complexity, and disorder. Lastly, in Chap. 10, Medina Ríos et al. review research that uses machine learning approaches to study how people perceive urban environments according to subjective dimensions like beauty and danger. Then, with a specific focus on Global South cities, they present a study on perception of urban scenes by people and machines. They use their findings from this study to discuss implications for the design of systems that use crowd-sourced subjective labels for machine learning and inference on urban environments.

We have edited this book to appeal to undergraduate and graduate students, academic and industrial researchers, and practitioners who are broadly interested in cognitive underpinnings of subjective visual experiences, as well as computational approaches to modelling and predicting them. The authors of this book provide

overviews of the current state of the art in their respective fields of study; therefore, chapters are largely accessible to researchers who may not be familiar with either prevailing computational, and particularly machine learning, practice, or with research practice in cognitive science. As such, we believe that researchers from both worlds will have much to learn from these chapters.

We are indebted to all the authors for their contributions, and hope that readers of this book will enjoy reading the fruits of their hard work as much as we have. Finally, we thank our editor, Springer, who gave us the opportunity to bring this project to life.

Bucharest, Romania — Bogdan Ionescu

Chicago, IL, USA — Wilma A. Bainbridge

Meylan, France — Naila Murray

Contents

The Ingredients of Scenes that Affect Object Search and Perception

Tim Lauer and Melissa L.-H. Võ

1 Introduction

What determines where we attend and what we perceive in a visually rich environment? Since we typically cannot process everything that is in our field of view at once, certain information needs to be selected for further processing. Models of attentional control often distinguish two aspects: Bottom-up attention (sometimes referred to as "exogenous attention") focuses on stimulus characteristics that may stand out to us, while top-down (or "endogenous") attention focuses on goal-driven influences and knowledge of the observer (e.g., Henderson et al., 2009; Itti & Koch, 2001). In this chapter, we focus on top-down guidance of attention and object perception in scene context; particularly, on top-down guidance that is rooted in generic scene knowledge—or *scene grammar* as we will elaborate on later—and is abstracted away from specific encounters with a scene, but stored in long-term memory.

Suppose that you are looking for cutlery in a rented accommodation. You would probably search in the kitchen or in the living room but certainly not in the bathroom. Once in the kitchen, you would probably readily direct your attention to the cabinets—it would not be worthwhile to inspect the fridge or the oven. Despite having a specific goal, certain items may attract your attention, such as a bowl of fruits or colorful flowers on the kitchen counter. If you found forks, you might expect to find the knives close by. While viewing the kitchen, you would probably not have a hard time recognizing various kitchen utensils, even if they were visually small, occluded or otherwise difficult to identify. In this example, one benefits from context information, from prior experience with kitchens of all sorts. That is, in the real

T. Lauer (✉) · M. L.-H. Võ
Goethe University Frankfurt, Frankfurt am Main, Germany
e-mail: tlauer@psych.uni-frankfurt.de

B. Ionescu et al. (eds.), *Human Perception of Visual Information*,
https://doi.org/10.1007/978-3-030-81465-6_1

world, objects are hardly ever seen in isolation but typically in similar, repeating surroundings which allows us to make near-optimal predictions in perception and goal-directed behavior (Bar, 2004; Oliva & Torralba, 2007; Võ et al., 2019). Figure 1 provides an illustration: While it is difficult to recognize the isolated object in the left panel, the availability of scene context (right panel) probably helps in determining the identity of the object (here an electric water kettle).

In this chapter, we will first review how attention is allocated in the real world from a stimulus-driven perspective. We will then outline important aspects of attentional guidance during visual search, followed by a section on contextual influences on object recognition—an integral part of search. In particular, we focus on what types of contextual information or "ingredients" the visual system utilizes for object search and recognition, a question that has remained largely unexplored until recently. To this end, we refer to diverse methodologies (like psychophysics, eye tracking, neurophysiology, and computational modelling) used at different degrees of realism (ranging from on-screen experiments, via virtual reality to studies in the real world). Finally, we will bring the findings together, discussing the relative contributions of various context ingredients to object search and recognition, as well as future directions and mutual benefits of human and computer vision research.

2 Attentional Allocation in Real-World Scenes

2.1 The Role of Low-Level Features

The bowl of fruits in our introductory example (see Fig. 1) would be expected based on the semantic scene context, but might initially stand out to us in terms of low-level features (e.g., color) that differ from the surroundings (e.g., white kitchen

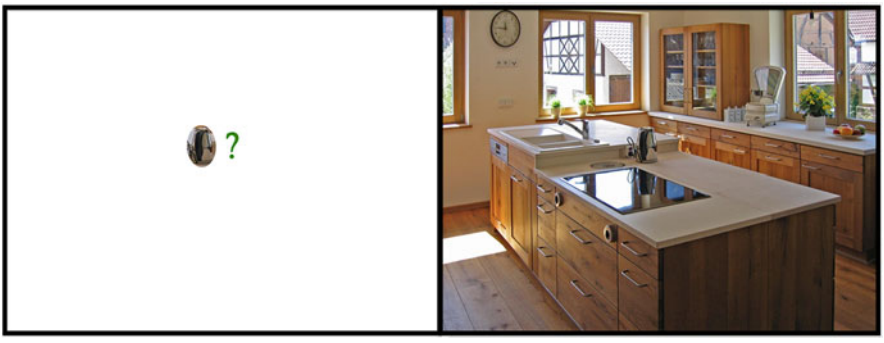

Fig. 1 While it is difficult to recognize the isolated object in the left panel, the kitchen context (right panel) may help in determining that the object is an electric water kettle. The kitchen scene was reproduced and adapted with permission from *Lignum Moebel*, Germany (https://lignum-moebel.de)

counter). Over the last two decades, several computational models of bottom-up, stimulus-driven attention have been put forth (for reviews, see Borji, 2019; Borji & Itti, 2013; Krasovskaya & Macinnes, 2019). A seminal early model of attention that inspired numerous other models is the saliency model by Itti and Koch (2000, 2001). Visual salience is defined as the "distinct subjective perceptual quality which makes some items in the world stand out from their neighbors and immediately grab our attention" (Itti, 2007). The model computes a salience map with regions that are likely attended by the observer based on low-level feature contrast (in intensity, orientation, and color) across spatial scales, motivated by receptive fields in the human visual system. Note that, as a proxy for *overt* visual attention, researchers often measure fixations and compare the empirical distributions to model predictions. However, visual attention is in principle not limited to the point of fixation and can be directed to regions outside of the fovea (commonly referred to as *covert* attention). Low-level saliency models have been shown to predict overt attention above chance under free viewing conditions (i.e., in the absence of a specific task), with highest predictability found for the first fixation (e.g., Parkhurst et al., 2002). Interestingly, these models capture where we direct our gaze merely based on low-level feature contrast, that is, without knowledge of image content or meaning (e.g., it is not known that the salient spot in the kitchen is a bowl of fruits or flowers).

2.2 The Role of Mid-Level Features and Objects

While low-level image features certainly play a decisive role for attentional allocation, it has been questioned whether attention is effectively attracted by such low-level features or rather higher-level features or objects that are not incorporated in low-level salience models (Einhäuser et al., 2008; Nuthmann & Henderson, 2010; Pajak & Nuthmann, 2013; Stoll et al., 2015). Objects often occur in locations that are salient (Spain & Perona, 2011)—oftentimes they make locations salient in the first place—and might thus be the driving force in attentional deployment (Schütt et al., 2019). Stoll et al. (2015) found that a state-of-the-art model of low-level salience and an object model predicted fixations equally well; however, when salience was reduced in regions that were relevant in terms of object content, the object model outperformed the salience model. Nuthmann and Einhäuser (2015) introduced a novel approach to investigate which image features influence gaze: Using mixed-effects models, they showed that mid-level features (e.g., edge density) and higher-level features (e.g., image clutter and segmentation) had a distinct contribution in gaze prediction as opposed to low-level features. Thus, many recent models incorporate mid to higher-level features in addition to low-level features to better predict fixation distributions in scene perception. To this end, deep neural networks (DNNs) have become increasingly popular and achieve benchmark performance in gaze prediction nowadays (Borji, 2019). One of the currently best-

performing networks, *DeepGazeII,* utilizes high-level features from a DNN trained on object recognition (Kümmerer et al., 2016).

2.3 The Role of Meaning

The role of scene meaning (or semantic informativeness) in attentional deployment while viewing real-world scenes has been studied for decades, and was recently systematically assessed by Henderson and colleagues (Henderson et al., 2018, 2019; Peacock et al., 2019a, 2019b). For a large number of local scene patches derived from scene images, they collected ratings of meaningfulness based on how informative or recognizable the patches were to observers. The authors then generated meaning maps which represent the spatial distribution of semantic features across a scene, comparable to a salience map (though not rooted in image-computable features). Meaning was shown to predict gaze successfully, as was low-level salience, but salience did not have a unique contribution when controlling for its correlation with meaning (Henderson & Hayes, 2017). This finding was replicated when predicting fixation durations instead of fixation distributions (Henderson & Hayes, 2018), and held across different tasks (Henderson et al., 2018; Rehrig et al., 2020), even when low-level image salience was highly task-relevant and meaning was not (Peacock et al., 2019a). However, it has been argued that the success of the meaning maps approach could be due to high-level image features that are not captured in classic salience models and could have strongly influenced observer's ratings of meaningfulness: *DeepGazeII,* which incorporates high-level object features, is able to outperform meaning maps at predicting fixations (Pedziwiatr et al., 2019).

Further, deriving meaning from objects in scenes has been shown to guide attention such that gaze tends to transition from one object to another object if the items are semantically related (Hwang et al., 2011; Wu et al., 2014a; for a review, see Wu et al., 2014b; see also De Groot et al., 2016). Objects that violate the global meaning of a scene (e.g., a mixer in the bathroom) strongly engage attention; they are typically looked at longer and more often than consistent objects (e.g., Cornelissen & Võ, 2017; De Graef et al., 1990; Friedman, 1979; Henderson et al., 1999; Loftus & Mackworth, 1978; Võ & Henderson, 2009b). While it has been established that attention can be "stuck" on these inconsistencies once they are spotted—even when they are irrelevant to one's current goals (Cornelissen & Võ, 2017, p.1)—it is a matter of debate whether they attract attention before they are fixated. Some studies have found semantic inconsistencies to influence initial eye-movements (e.g., the critical object is fixated earlier than a consistent object) (Becker et al., 2007; Bonitz & Gordon, 2008; Coco et al., 2019; Loftus & Mackworth, 1978; Nuthmann et al., 2019; Underwood et al., 2007, 2008; Underwood & Foulsham, 2006), yet other studies did not find indication for attention capture by inconsistencies (Cornelissen & Võ, 2017; De Graef et al., 1990; Furtak et al., 2020; Henderson et al., 1999; Võ & Henderson, 2009b, 2011). These mixed results may be related to characteristics of the scene stimuli (e.g., line drawings, photographs, or 3D-rendered scenes with

varying degrees of clutter) and/or more or less controlled characteristics of the critical objects (e.g., size, eccentricity, salience).

With the rise of fully labeled image databases like LabelMe (Russell et al., 2008) assessing the semantic relatedness between objects and their scene contexts as well as inter-object relatedness has become easier. For instance, using graph theory by treating objects as nodes and assigning different weights to their connections has provided new avenues to determine clusters of semantically related objects within scenes—which we have started to call "phrases"—or prominent objects therein that anchor predictions about the location and identity of other objects nearby (for more details, see Sect. 4.3; Boettcher et al., 2018; for reviews, see Võ, 2021; Võ et al., 2019). Objects that do not fit their context tend to be regarded as surprising or interesting and can affect where we attend to in scenes.

2.4 The Role of Interestingness and Surprise

While the role of image features has been studied extensively (for reviews, see Borji, 2019; Borji & Itti, 2013; Krasovskaya & Macinnes, 2019), relatively little is known about how other factors such as interestingness or surprise modulate attentional deployment. Elazary and Itti (2008) proposed that interesting objects are in fact visually salient: Observers who contributed to the LabelMe database—a large collection of scenes with object annotations (Russell et al., 2008)—tended to label those objects that were salient even though they were free to choose which objects to label. In another study, when explicitly asked which scene locations are interesting, the choice of locations was largely similar across observers and correlated with fixation distributions of other observers (Masciocchi et al., 2009). Behavioral judgements and eye movements were also correlated with predictions of a salience model, yet not as highly as one would expect if salience was the only driving factor of interestingness. The authors concluded that there are both bottom-up and top-down influences on what we perceive as interesting and where we attend in an image (see also Borji et al., 2013; Onat et al., 2014). Other studies have shown that, beyond an influence of low-level salience, attentional allocation is modulated by the affective-motivational impact of objects or their importance for the scene ('t Hart et al., 2013; Schomaker et al., 2017), and that attention is attracted by surprising image locations in a Bayesian framework (e.g., Itti & Baldi, 2005). Moreover, some types of objects hold a special status: Text and faces, for instance, have been shown to greatly attract attention in scenes (see Wu et al., 2014b).

Taken together, inspired by early models of low-level salience, more recent research highlights the importance of higher-level features and indicates that attention in scenes is largely object-based—with some objects attracting and/or engaging attention more than others. While DNNs achieve benchmark performance in a variety of tasks nowadays and have become increasingly popular in fixation prediction, more research is needed to see how they will further our understanding of human attention mechanisms. Further, it will be crucial to shed more light on

when during scene viewing various features exert influence on attentional allocation. Schütt et al. (2019) disentangled the contribution of low and higher-level features to fixation distributions over time, showing that the influence of low-level features is mostly limited to the first fixation and that higher-level features, as incorporated in *DeepGazeII*, predict fixations better starting 200 ms after stimulus onset. Despite the popularity of DNNs, a shortcoming of data-driven approaches is that they do not capture some aspects of human visual attention such as singleton (or "odd one out") detection in artificial stimuli (even when the training data is adjusted, e.g., Kotseruba et al., 2020).

3 Guidance of Attention during Real-World Search

While the processing of image features can certainly play a role in where we attend, especially when free-viewing scenes, we are rarely ever mindlessly looking around. Instead, we tend to be driven by various agendas and task demands, one of which is the need to locate something or somebody. The interplay of bottom-up image features and more cognitively based, top-down influences during search is complex. As Henderson (2007) put it: "In a sense, we can think of fixation as either being "pulled" to a particular scene location by the visual properties at that location, or "pushed" to a particular location by cognitive factors related to what we know and what we are trying to accomplish" (p. 219). However, it should be noted that it is not always straightforward to strictly delineate between bottom-up and top-down influences (Awh et al., 2012; see also Teufel & Fletcher, 2020); we are certainly not claiming that the aspects presented here are one *or* the other.

Traditionally, visual search was studied using simple artificial displays of randomly arranged targets and distractors (e.g., "find the letter T among several instances of the letter L"). The main measure was—and still is— reaction time (RT) as a function of set size (i.e., the number of items in the display). With increasing set size, RT is consistently longer in such a task, in equal steps, indicating that attention is serially deployed to one item after another (see Wolfe, 2020; Wolfe & Horowitz, 2017). However, in some cases, it is not necessary to inspect all items in the display: In "classic guided search" theory, a limited set of target features (e.g., color, motion, orientation, size) can guide attention in a top-down manner, narrowing down the number of possible items (for reviews, see Wolfe, 2020; Wolfe et al., 2011b; Wolfe & Horowitz, 2017). For instance, when looking for a red "T" among some red and some black "L"s one can disregard all black items. To this end, "feature binding" takes place: The shape and the color of the target are bound together in order to reject distractors as well as recognize the target(s). While the field has learned a lot from these types of experiments that mostly used meaningless stimuli, search in real-world scenes seems to be strongly influenced by other guiding factors.

Scenes are not random assemblies of features but most often structured and meaningful, which allows us to perform searches with remarkable efficiency. For instance, when looking for a teddy in the bedroom, fixations tend to cluster around

the bed even if the target is not present and cannot guide attention by means of its features (see Võ et al., 2019). Search for objects in scenes appears to be much more efficient than search for isolated objects in random arrays, although it can be challenging to define a scene's set size adequately (see Wolfe et al., 2011a). As proposed in the *cognitive relevance framework,* search in scenes is mainly guided by cognitive factors such as prior knowledge and current goals (Henderson et al., 2009; for a review, see Wolfe et al., 2011b).

What makes search in the real world so efficient despite the wealth and complexity of information contained in the visual input? While no one would doubt that scene context aids object search, relatively little is known about which "ingredients" of real-world scenes effectively guide attention, what their relative contributions are, and when they contribute during the search. In the following, we attempt to shed more light on these ingredients.

3.1 The Role of Scene Gist

One line of work addressed the question of whether an initial brief glance at a scene influences attentional allocation. Within a fraction of a second, observers can obtain the "gist" of a scene, a coarse representation of its spatial properties and meaning that does not require the selection of individual objects (Greene & Oliva, 2009a, 2009b; Rousselet et al., 2005). While there is no universal account of scene gist, many definitions (including ours), state that gist allows the categorization of scenes at a basic level. For instance, one may categorize a scene as a kitchen and tell that it comprises something like a kitchen counter but not yet grasp that there are a toaster and a mixer resting on any of the surfaces. That is, one may "see the forest without representing the trees" (Greene & Oliva, 2009a). A brief glance in the range of milliseconds is too short to make a saccade and thus to foveate selected parts of the scene in order to perceive them with fine detail. In fact, scene gist recognition does not depend on the high visual acuity of the fovea; it can be achieved even when the scene is blurred or when only peripheral information is available (e.g., Loschky et al., 2019). One fundamental aspect of scene gist is spatial layout information. As demonstrated in the *spatial envelope model* and supported by behavioral studies, scenes can be categorized based on their global properties, such as the global shape, without the need to identify any objects in the scene (Oliva & Torralba, 2001, 2006). This way of processing the scene is considered to be largely feed-forward and, in terms of search guidance, is assumed to take place on a "nonselective pathway" that parallels a "selective pathway" which binds features and recognizes individual objects (Wolfe et al., 2011b). Note that objects can also be an important source of information for scene categorization (MacEvoy & Epstein, 2011), especially for indoor scenes that are not always easily distinguishable in terms of their global properties.

To investigate how a brief glance at a scene guides search behavior, researchers have used the flash-preview moving window paradigm (Castelhano & Henderson,

2007; Võ & Henderson, 2010, 2011; Võ & Schneider, 2010; Võ & Wolfe, 2015): It initiates with a brief preview of a scene, followed by a target word and a search phase in which observers look for the target object in the original scene but through a gaze-contingent window that only reveals a small area of the scene at the current point of fixation. Given that the scene as a whole is not perceived during the search phase, this paradigm allows experimenters to assess the contribution of the scene's initial global percept to visual search. Note, however, that this contribution may be weaker under more natural search conditions in which the entire scene can be processed online during the search as well (see Võ & Wolfe, 2015). A scene's preview has been shown to influence visual search consistently in these studies, even when it was as short as 50 ms (Võ & Henderson, 2010). Võ and Schneider (2010) manipulated the type of context information that was available in the scene preview, selectively preserving either the global scene background or local objects (for an illustration, see Fig. 2). The availability of the scene background, conveying the spatial layout of the scene, resulted in faster detection of the targets and required fewer fixations compared to a control condition, whereas a preview of local objects did not facilitate search. Thus, a coarse representation of a scene's structure and meaning appears to already guide visual search effectively. Interestingly, knowing only the category of the scene does not seem to be sufficient, as was shown when a searched scene was

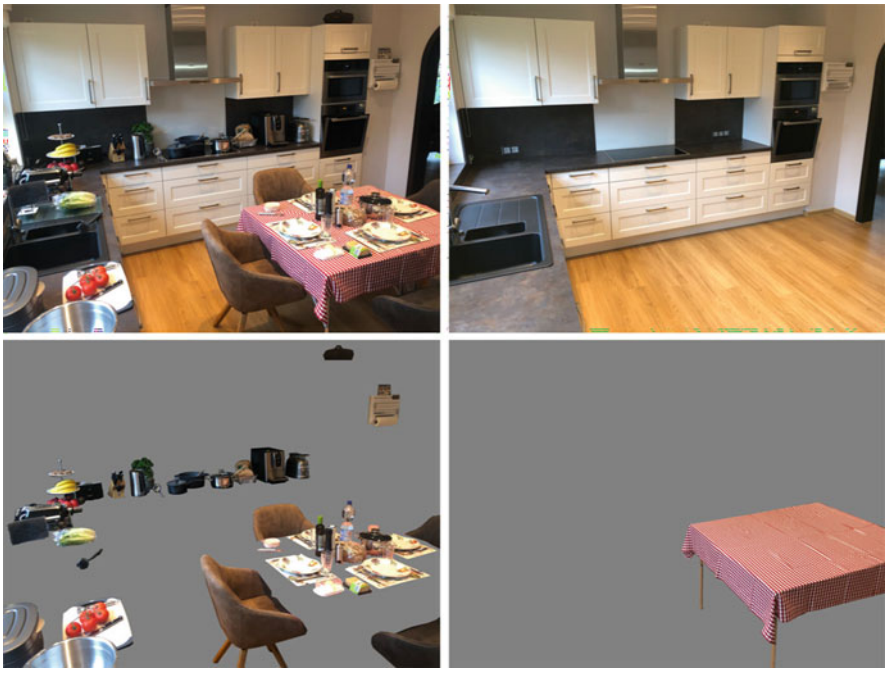

Fig. 2 Illustration of a kitchen scene (top left) that can be divided into the background (top right), local objects (bottom left) as well as an anchor object (bottom right)

primed by a different scene exemplar from the same category or by a word label of the category. Yet, a scene that is semantically inconsistent with a target (e.g., a mug of paint brushes in a bedroom) can facilitate search given that the object occurs in a reasonable location (Castelhano & Heaven, 2011, for a review see Castelhano & Krzyś, 2020).

The spatial layout of a scene can provide us with important constraints regarding the location of objects. For example, the occurrence of objects is constrained by the laws of physics such that objects rest on surfaces rather than hovering in the air. Even when we do not fully grasp a scene's meaning, we may be able to tell where its major surfaces lie (e.g., kitchen counters, tables, etc.) (see Fig. 2) and/or where the sky and the horizon are located. Moreover, two objects usually do not occupy the same physical space (Biederman et al., 1982), and we know where certain objects typically occur (e.g., a rug is often located on the floor) (Kaiser & Cichy, 2018; Neider & Zelinsky, 2006). Incorporating likely vertical object locations in a low-level salience model can significantly improve gaze prediction, as was demonstrated in the *contextual guidance model* (see Oliva & Torralba, 2006). More recently, the *surface guidance framework* was introduced, proposing that attention is allocated to surfaces in the scene that are related to the target object (Castelhano & Heaven, 2011; Pereira & Castelhano, 2014, 2019; for a review, see Castelhano & Krzyś, 2020).

3.2 The Role of Local Objects

Another line of work investigated the influence that selected parts of the scene, specifically objects, have on attentional allocation. In a naturalistic search task, Mack and Eckstein (2011) instructed participants to search for objects on tables while wearing mobile eye tracking glasses. The target object (e.g., a fork) was either located near a so-called cue object with which it would likely co-occur in natural scenes (e.g., a plate) or elsewhere (close to other objects). Targets were found faster if they were located near cue objects, and cue objects were fixated more frequently than other objects surrounding the targets, suggesting that object co-occurrence in the real world can boost search performance. In another study, in which participants inspected scene images or searched for targets therein, the LabelMe database of scenes with object annotations was used to determine the semantic relatedness of the currently fixated object to other objects in the scene or to the search target (Hwang et al., 2011). Gaze was shown to transition more likely to objects that are semantically related to the currently fixated object, even when the objects were not in close proximity. Moreover, the search data revealed that the influence of target-based semantic guidance increased throughout the trial. The finding of likely transitioning between related objects was replicated even when the objects were cropped (removed) from the scenes but not when discarding spatial dependencies among the cropped objects by re-arranging them (Wu et al., 2014a). When a preview of the original scene was added in order to provide gist information,

there was no indication of increased semantic guidance. Moreover, there is evidence that the functional arrangement of objects influences gaze direction in the absence of scene context (e.g., a key that is arranged such that it can or cannot be inserted in a lock) (Clement et al., 2019). In object arrays, semantic information can be extracted extrafoveally and can guide even the first eye movement during search (Nuthmann et al., 2019). Taken together, both the semantic relation of objects as well as their spatial dependencies appear to be relevant for attentional allocation during search.

3.3 The Role of Anchor Objects

There seem to be certain objects that predict not only the occurrence, but particularly the location of other objects within a scene. Boettcher et al. (2018) explored the role of spatial predictions in object-based search guidance, introducing the concept of *anchor* objects. Anchors are typically large, static objects (i.e., they are rarely moved) that give rise to strong predictions regarding the identity and location of *local* objects clustering around them (e.g., the table may predict the position of a chair, a glass of water, and the salt). By contrast, local objects do not necessarily predict the location of other local objects (e.g., when searching for the salt, the location of a glass might not be that informative) (see Fig. 2). Using the LabelMe database, the concept of anchor objects was operationalized through four factors: variance of spatial location, frequency of co-occurrence, object-to-object distance, and clustering of objects (see Boettcher et al., 2018; c.f. Võ et al., 2019). In a series of eye tracking experiments, observers searched for target objects in images of 3D-rendered scenes (e.g., bathroom) that were manipulated to either contain a target-relevant anchor (e.g., shower) or a substitute object that was chosen to also be semantically consistent with the scene and of similar size (e.g., cabinet). Compared to the substitute objects, relevant anchors affected search performance such that there was a reduction in reaction time, scene coverage, and the time to transition from the anchor to the target. In line with this, in a recent virtual reality experiment, participants were slower at locating target objects when anchors were concealed by grey cuboids of similar dimensions compared to when they were fully visible (Helbing et al., 2020). Randomly re-arranging the anchors (or cuboids) resulted in an opposite effect, that is, targets were located faster in the cuboid condition, suggesting that both the identity and spatial predictions of anchors are crucial for their ability to guide search. Note that these inherent spatial predictions distinguish anchor objects from the notion of diagnostic objects (e.g., MacEvoy & Epstein, 2011) which may be important for conveying scene meaning and facilitating scene categorization, but need not yield precise predictions of the occurrence of other objects (Võ et al., 2019). It seems likely that anchor objects can be identified even in the periphery (see Koehler & Eckstein, 2017b, for a demonstration of peripheral extraction of object cues) and thus they might provide an effective way to locate smaller targets, building a bridge between the global scene and local objects.