Chapter 8
Bottom-Up Processing in Complex Scenes: A Unifying Perspective on Segmentation, Fixation Saliency, Candidate Regions, Base-Detail Decomposition, and Image Enhancement

Boyan Bonev and Alan L. Yuille

Abstract Early visual processing should offer efficient bottom-up mechanisms aiming to simplify visual information, enhance it, and direct attention to make high-level processing more efficient. Based on these considerations, we propose a unified approach which addresses a set of fundamental early visual processes: segmentation, candidate regions, base-detail decomposition, image enhancement, and saliency for fixations prediction. We argue that for complex scenes all these processes require hierarchical segmentwise processing. Furthermore, we argue that some of these visual tasks require the ability to decompose the appearance of the segments into "base" appearance and "detail" appearance. An important, and surprising, result of this decomposition is a novel method for successfully predicting human eye fixations. Our hypothesis is that we fixate on segments that are not easy to model, e.g., are small but have a lot of detail, in order to obtain a higher resolution representation for further analysis. We show performances on psychophysics data on the Pascal VOC dataset, whose images are non-iconic and particularly difficult for the state-of-the-art saliency algorithms.

Keywords Bottom-up visual processing • Image segmentation • Base-detail decomposition • Saliency

8.1 Introduction

Low-level vision is visual processing that treats images as patterns and makes no specific assumptions about the objects that might be present or the structure of the scene. In short, the processing is generic and intended to be suitable for all images, regardless of their semantic content or high level layout. Examples of

B. Bonev (✉) • A.L. Yuille
Department of Statistics, University of California, Los Angeles, CA, USA
e-mail: bonev@ucla.edu; yuille@stat.ucla.edu

S.-W. Lee et al. (eds.), *Recent Progress in Brain and Cognitive Engineering*,
Trends in Augmentation of Human Performance, DOI 10.1007/978-94-017-7239-6_8

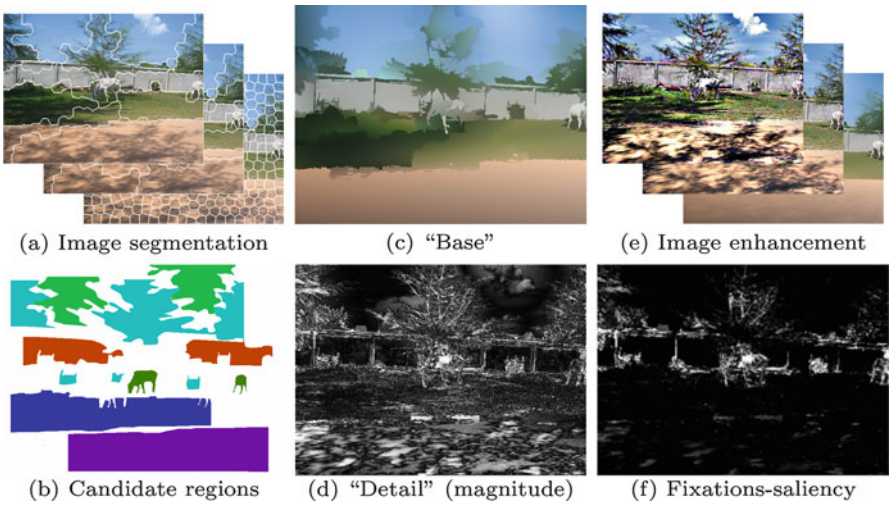

Fig. 8.1 We propose a unified approach for several low-level visual processes: (**a**) image segmentation – a hierarchy of image partitions at multiple levels; (**b**) candidate regions – a pool of possibly overlapping proposals for further study by object recognition methods (best candidates illustrated); (**c**,**d**) "base-detail" decomposition – expressing the image as the sum of a non-local smooth appearance term and s residual, or detail, which captures the texture patterns; (**e**) image enhancement – controlling the amount of detail in the image; (**f**) saliency for fixations prediction – a model predicting bottom-up human visual attention

low-level vision tasks include segmentation, candidate regions or object proposals, and image enhancement. Low-level processing is typically performed in preparation for high-level tasks, and is used to allocate computational resources for more detailed processing. In mammalian visual systems low-level vision is believed to be performed in the retina and area V1 of the visual cortex.

In this paper we propose a unified framework for several low-level vision tasks (Fig. 8.1) that are typically modeled separately. These tasks include the generation of hierarchical image segmentations, proposing candidate regions for object detection and recognition, base-detail decomposition – where an image is decomposed into a visual summary plus fine details – image enhancement and the prediction of human eye fixations.

We start by producing a hierarchical decomposition of the image into segments which have roughly uniform homogeneity as measured by texture and color cues. Segments at higher levels of the hierarchy are generally larger and less homogeneous. But in our approach, it is important that the size of segments within each level of the hierarchy have different sizes because some image regions (e.g., sky) are much more homogeneous than others (e.g., a road containing several cars).

Different segments of the hierarchy are combined into groups of up to three to make proposals for the positions and shapes of objects and background "stuff" [10], which we refer to as candidate regions. They consist of a pool of 500–1500 regions which are later evaluated by a high-level method, which is out of the scope of this

paper. The high-level method computes category-specific scores to identify regions which correspond to object or background categories.

We define "base-detail" decomposition as the separation of the image into a coarse description of the image appearance, and a description containing the texture and details. The image is the sum of both. More precisely, the base is obtained by fitting smooth appearance models (polynomials) to the image segments and the detail is the residual. For examples, see Figs. 8.1c, d and 8.8. This base-detail decomposition enables us to process the image in several ways, such as enhancing the details and/or the base. For example, we can remove the shadows (details) from a grass lawn (base). Surprisingly, as we now discuss, we can use base-detail decomposition to predict human eye fixations for free viewing.

It is well known that when humans examine an image they do not gaze on it uniformly but instead they fixate on certain parts of the image. The fixation saliency model we propose favors small segments which have strong details. This has the following intuition: large segments are typically homogeneous regions (e.g., sky, water, or grass) which may be easily processed (i.e., classifying these regions may be easy using methods which use summary image statistics and do not model the detailed spatial relations). The detail is less important in the large segments but in small segments the detail may correspond to structures which require more detailed models to process. We describe experiments showing that our fixation saliency model predicts human fixations with a state-of-the-art performance on complex datasets, like Pascal [17] and Judd [31].

Our work is motivated both by attempts to understand how primate visual systems work and by efforts to design computer vision systems with similar abilities. We provide a computational model for performing these visual tasks but in this paper we do not develop any detailed biological evidence for this theory. Instead we concentrate on performance on complex visual scenes, instead of artificial stimuli, because we think it is important to model visual abilities in real-world conditions.

8.2 Background and Related Literature

There is an enormous literature on segmentation much of it using Markov Random Field (MRF) models [22]. Our work follows the alternative strategy of decomposing images into subregions which have roughly similar statistical image properties [1, 33, 45, 52]. There are a variety of hierarchical approaches which exploit the intuition that image structures occur at different scales and that multi-scale is required to capture long-range interactions within the images [19, 53]. Our approach to hierarchies follows the strategy of starting with an over-segmentation of the image, produced by an efficient algorithm like [1], followed by recursive grouping to get larger segments at different levels of the hierarchy [4]. This relates closely to Segmentation by Weighted Aggregation [20], a recent variant [3], and extensions to video segmentation [48].

Detecting candidate regions, which make proposals for the positions and sizes of objects, is a new but increasingly important topic in computer vision. This is because it offers an efficient way to apply powerful methods, such as Deep Convolutional Neural Networks (DCNN) [34], to detect and recognize objects in images. Instead of needing to apply DCNNs exhaustively, at every image position and scale, it is only necessary to apply them to a limited number of candidate regions. Our method for detecting candidate regions differs from existing methods because we propose regions for both objects and background regions or "stuff" (e.g., sky). Recent work on detecting candidate regions includes methods which group segments into combinations [5, 6]. Most methods in the literature have been evaluated for finding segments which cover foreground objects [46], while ours detects background classes as well. Finally, there are other methods which differ in that they mainly exploit the edges instead of the appearance statistics [15, 29, 55]. We should also mention hierarchical segmentation which has been used to learn models of objects [43].

There is no existing work that directly addresses "base-detail" segmentation, but there is a large literature on closely related topics. In the digital image processing community there is a related concept, "base-detail separation", but it is performed locally [7] by applying bilateral filters. A related topic is gain control which has been studied in primate visual systems, particularly in the retina, and seeks to compress the dynamic range of the input intensity while preserving the local contrast and detail [14, 42]. We note that detection of detail is also at the heart of many super-resolution methods [54] and it is related to image enhancement. Enhancement approaches do not typically use segment-based methods [25, 41] and instead use local methods like the bilateral filter in [7] or the weighted least squares [18]. There are some exceptions, like [49] where segment-wise exposure correction is proposed.

Another related topic is work in the shape from shading community, where intensity patterns are decomposed into smoothly varying shading patterns and more variable texture/albedo components [9, 26, 27] (here the base roughly corresponds to the shading and the texture/albedo to the detail). Researchers in shape from shading make prior assumptions for performing the decomposition into shading and texture/albedo [8] (which are not needed if the same object is viewed under different lighting conditions [47]). Similar decompositions assumptions are also applied to the classic Mondrian problem [32].

Predicting human eye fixations is a long studied research topic [30]. In this paper we address only bottom-up saliency prediction, as performed in a free-viewing task, and do not consider top-down processes involving which involve cognitive factors, e.g., eye fixations when performing a task such as counting the animals in an image. One of the first successful methods for predicting human eye fixations was Itti's original model [30]. Image signature is a simple method which give good results [28] and other recent methods are reviewed in [11]. The most successful current method is Adaptive Whitening Saliency [21] and we make comparisons to it in our experiments. Finally, there are other works [37, 38] which studies the saliency of visual objects and use candidate regions to make predictions [13]. Objects are judged to be salient based on the number of eye fixations which occur within them.

By contrast, fixation saliency only predicts positions and outputs a fixations map (in Sect. 8.4.3, see Fig. 8.16, second column). The eye fixation saliency models we propose is based on base-detail decomposition, which makes it substantially different from any method in the literature. Our experiments show it performs at the state of the art.

Finally, although biology is out of the scope of this paper, we find it interesting that recent biological vision studies suggest that early visual processing is more sophisticated than traditional models of the retina and V1, which mainly emphasize linear spatiotemporal filters. For example, studies of the retina suggest that it is "smarter than scientists believed" [23] and contains a range of non-linear mechanisms which might perhaps be able to implement parts of the theory of theory we propose here. Moreover, there is growing appreciation of the richness of computations that can be performed in area V1 of the visual cortex, including possibly fixation saliency [51].

8.3 Method

In this section, we describe the details of the proposed approach. We address a set of fundamental low-level vision processes: segmentation, candidate regions and salient objects proposals, base-detail decomposition, image enhancement, and bottom-up saliency. Instead of being treated as separate tasks, we address them in terms of a unified approach of bottom-up vision processing.

8.3.1 Segmentation: Hierarchical Image Partitioning

Image segmentation is a classic task of low-level vision. But in this paper we do not consider segmentation as a goal in itself. Instead, we seek to obtain a hierarchy of segmentations, or partitions of the image into segments, which can be used as components for other processing, as will be described in the next subsections.

An image partition is a decomposition of the image into non-overlapping subregions, or *segments*. More formally, we decompose the image lattice $\mathcal{D}$ into a set of segments $\{\mathcal{D}_i : i = 1, \ldots, n\}$ such that:

$$\mathcal{D} = \bigcup_{i=1}^{n} \mathcal{D}_i, \quad \text{s.t. } \mathcal{D}_i \bigcap \mathcal{D}_j = \emptyset, \ \forall i \neq j.$$

A hierarchical partition of an image is a set of decompositions indexed by hierarchy level $h = 1, \ldots, H$. Each level gives an image partition $\mathcal{D} = \bigcup_{i=1}^{n_h} \mathcal{D}_i^h$, where n_h is the number of segments in the partition at level h. The decompositions are *nested* so that a segment $\mathcal{D}_i^h$ at the hierarchy level h is the union of a subset of segments

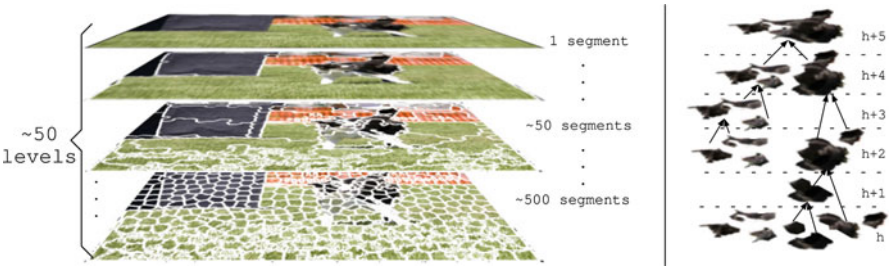

Fig. 8.2 *Left*: Multiple levels in a hierarchy. Segments with a good coverage of objects or parts may happen at different levels. 80 % to 90 % of the segments can be discarded because they go across boundaries of objects or because they don't cover a large area of an object. *Right*: Segments at level $h+1$ are composed of one or two segments in level h

at the previous level $h-1$, so that $\mathcal{D}_i^h = \bigcup_{j \in Ch(\mathcal{D}_i^h)} \mathcal{D}_j^{h-1}$, where $Ch(\mathcal{D}_i^h)$ denotes the child segments of segment i at level h (in this paper each segment is constrained to have at most two immediate children, see Fig. 8.2-right). This enables us, by recursion, to express a segment in terms of compositions of its descendants in many different ways. In particular, we can decompose a segment into its descendants at the first level, $\mathcal{D}_i^h = \bigcup_{j \in Des(\mathcal{D}_i^h)} \mathcal{D}_j^1$. This hierarchical structure is common in the segmentation literature, for example in [4]. Figure 8.2 illustrates the hierarchical partitioning of an image.

In this paper, our hierarchical partitioning is designed based on the following related considerations. Firstly, we prefer segments to have roughly homogeneous image properties, or *statistics* $\vec{S}$ (e.g., color/texture/detail) at each level, which means that segments at the same level can vary greatly in size (e.g., segments on the grass in Fig. 8.2 will tend to be larger than in less homogeneous regions of the image, like the dog). Secondly, segments at higher levels should be less homogeneous because they are capturing larger image structures (e.g., by merging more homogeneous image structures together). Thirdly, segments are likely to have edges (i.e., image intensity discontinuities) near their boundaries. Fourthly, we want an efficient algorithm which can dynamically compute this hierarchy using local operations by merging/grouping segments at level $h-1$ to compose larger segments at level h.

Our work is guided by standard criteria for image segmentation [33, 45, 52] which propose minimizing a cost function of form:

$$E(\{\mathcal{D}_i\}, \{\vec{S}_i\}) = \sum_i \sum_{x \in \mathcal{D}_i} |\vec{S}_i - \vec{S}(x)|^2 - \lambda \sum_i \sum_{x \in \partial D_i} e(x). \tag{8.1}$$

Here $\vec{S}(x)$ denotes image statistics at position x (e.g., color, texture features), $\vec{S}_i$ is summary statistics of the region i, λ is a non-negative constant, and $e(x)$ is a measure of edge strength (taking large values at image discontinuities), and ∂D_i is the boundary of segment $\mathcal{D}_i$.

We initialize our algorithm by using the SLIC [1] algorithm to compute the lowest level, $h = 1$, of our hierarchy. Essentially, SLIC performs an expectation-minimization of (8.1) for a fixed number n_1 of segments. It uses the color and position as statistics, without including an edge term, that is, $\lambda = 0$ in (8.1). More precisely, $\vec{S}(x) = (l(x), a(x), b(x), x)$, where l, a, b specify color channels of the Lab color opponent space and x denotes 2D spatial position.

Next, we proceed to construct the hierarchy by grouping/merging segments which have similar image statistics. The statistics are extended to include texture, shape of segments, and the variance of color (we do not use these statistics at the bottom-level because the segments are too small to compute them reliably). More precisely, $\vec{S}$ is given by the mean and the standard deviation of the Lab color space components and the first and second derivatives of the l channel, $(l, a, b, \nabla_x l, \nabla_y l, \nabla_x^2 l, \nabla_y^2 l)$, the centroids of the segment and dimensions of its bounding box (c_x, c_y, d_w, d_h). When performing merging, we use an asymmetric criterion which requires comparing the difference between the statistics of the union of the two segments i and j, $\vec{S}_{i \bigcup j}$, and the statistics of its segments $\vec{S}_i, \vec{S}_j$, that is, $||\vec{S}_{i \bigcup j} - \vec{S}_i||$ and $||\vec{S}_{i, \bigcup j} - \vec{S}_j||$. This is because our segments are allowed to have different sizes and we want to discourage bigger segments from merging with smaller segments if this will change much the statistics of one of them. Intuitively, a big segment is likely to have little change on its statistics by merging to a small one, but we want to ensure that the small one does not undergo a big change in its statistics. At each level of the hierarchy we allow the top-ranked 30 % segments to merge to another segment (rank is based on asymmetric criterion described above and prioritizes similar segments) but prevent merges where the asymmetric condition is violated. Merging is allowed between 1st and 2nd neighbors only. The precise details are described in [10].

The output is a hierarchical partition of the image. It is expressed as a set of segments $\{\mathcal{D}_i^h\}$, $1 \leq h \leq H$, $1 \leq i \leq n_h$, where h is the hierarchy level. At the highest level, $n_H = 1$. Each image region $\mathcal{D}_i^h$ and has statistics $\vec{S}_i^h$. Each level h gives a partition of the image $\mathcal{D} = \bigcup_{i=1}^{n_h} \mathcal{D}_i^h$. Each segment is composed of a set of child segments, $\mathcal{D}_i^h = \bigcup_{j \in Ch(\mathcal{D}_i^h)} \mathcal{D}_j^{h-1}$. Each segment can also be associated to its descendant segments at the $h = 1$ level: $\mathcal{D}_i^h = \bigcup_{j \in Des(\mathcal{D}_i^h)} \mathcal{D}_j^1$. This hierarchical partition can be used directly for image segmentation but, in the spirit of this paper, we think of it as a representation that can be used to address several different visual tasks as we will describe in the next few sections.

8.3.2 Candidate Regions

This section shows how to use the hierarchical partition to obtain candidate regions, or proposals, for both foreground objects and background regions, or "stuff" (e.g., sky, water, grass). Proposing candidate regions enables algorithms to concentrate computational resources, e.g., deep networks, at a limited number of locations (and

sizes) in images (instead of having to search for objects at all positions and at all scales). It also relates to the study of *salient objects* [2, 37], where psychophysical studies show that humans have tendencies to look at salient objects [16]. Note that salient objects, however, do not predict human eye fixations well [12] and these can be better described by bottom-up saliency cues [30] in a free-viewing task. However, methods that combine bottom-up saliency cues with proposals for candidate regions do perform well for both predicting human eye fixations and for the detection of salient objects [38].

We create candidate region proposals by the following strategy. Firstly, we select a subset of *selected segments* from the hierarchical partition of the image. These segments are chosen to be roughly homogeneous but as large as possible. Secondly, we make compositions of up to three selected segments to form a candidate region. These compositions obey simple geometric constraints (proximity and similarity of size). The intuition for our approach is that many foreground objects and background "stuff", can be roughly modeled by three segments or less, see Fig. 8.4. This intuition was validated [10] using the extended labeling of Pascal VOC [40] which contained per-pixel labels of 57 objects and "stuff".

The selected segments are chosen by computing the entropy gain of the combination of two child segments into their parent segment. If the entropy gain is small, then we do not select the child segments because this is evidence that they are part of a larger entity. But if the entropy gain is large, then we add the child segments to our set of *selected segments*. More precisely, we establish a constant threshold G for the entropy gain g after merging two segments $\mathcal{D}_i^h, \mathcal{D}_j^h$ into their parent $\mathcal{D}_m^{h+1} = \mathcal{D}_i^h \bigcup \mathcal{D}_j^h$. The entropy gain is defined to be:

$$g = \mathcal{H}(\mathcal{D}_m^{h+1}) - \left\{\mathcal{H}(\mathcal{D}_i^h) + \mathcal{H}(\mathcal{D}_j^h)\right\}. \tag{8.2}$$

Here $\mathcal{H}(\mathcal{D}_i^h)$ is the entropy of a segment i at level h, computed from the statistics $\{\vec{S}_k^1\}$, $k \in Des(\mathcal{D}_m^h)$ of its descendant segments at level $h = 1$ (Fig. 8.3). The entropy is computed in a non-parametric manner [10] using the approximation proposed in [35]. See an example of triplets of selected-segments in Fig. 8.4.

Fig. 8.3 *Entropy gain* (Sect. 8.3.2): When segments **a** and **b** are merged, the increase of entropy is not as big as if they were merged with **c**. *Homogeneity criterion* (Sect. 8.3.3.1): Segment **c** is homogeneous. It presents smooth variation due to shading and lighting. Segments **a** and **b** are not homogeneous. Both entropy and homogeneity are calculated from the small (first level $\mathcal{D}_i^1$) segments, illustrated with white contours

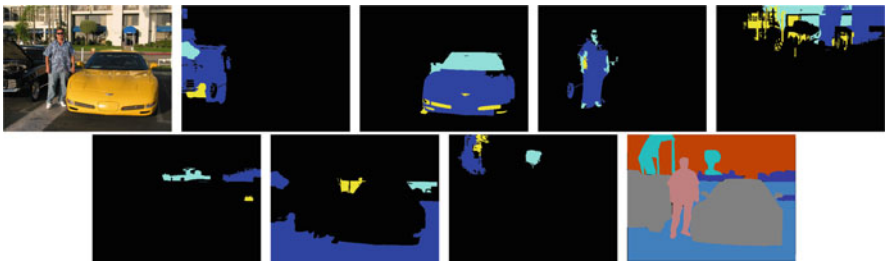

Fig. 8.4 Examples of candidate regions for foreground and background regions. Left-to-right and top-to-bottom: image, top three selected-segments for left car, right car, person, building, grass, ground, trees, and ground truth. Most objects are covered well by two to three selected-segments

8.3.3 Base-Detail Decomposition

This section analyzes the image intensities within the segments by decomposing the image into base and detail. The base $B(x)$ component is the approximate color of the region, and is required to be spatially smooth. The detail $R(x)$ is the residual $R(x) = I(x) - B(x)$ and can contain general texture, such as the patterns of grass on a lawn, or structured detail such as the writing on the label of a wine bottle.

Base-detail relates to several well studied phenomena. Firstly, it is similar to the task of preserving image contrast (i.e. the detail) performed by the early visual system when doing gain control. Secondly, it relates to the decomposition $I(x) = a(x) \cdot \vec{n}(x) \cdot \vec{s}(x)$ of images into albedo, normals and illumination when computing intrinsic images or the 2.5D sketch. But this decomposition is higher-level, relying on concepts like geometry and lighting sources, while we are modeling at a lower level. We note that in some special situations the base and the detail of a segment may correspond to the shading and the albedo of an object. Thirdly, base-detail also relates to transparency – e.g., the viewing of images through a dirty window (the dirt is the detail) – or when there is partial occlusion like tree leaves in front of a building (leaves are details). More generally, within image regions there is base appearance which changes smoothly within segments and detail which changes in a more jagged manner. This differs from the base-detail separation [7] studies in the image processing literature, which is obtained by local smoothing methods and not in a segment-wise manner.

We address base-detail decomposition in two steps. Firstly, we seek a segmentation of the image into regions which are as homogeneous and as large as possible. This is done by selecting a subset of those hierarchy segments $\{\mathcal{D}_i^h\}$ which are *maximally large and homogeneous* and form a partition of the image. Note that this includes segments at different levels h of the hierarchy. Secondly, within each segment we fit a low-order polynomial to the color intensities and define the best fit polynomial to be the base (see Sect. 8.3.3.2). We obtain the detail by computing the residual between the image and the base.

8.3.3.1 Finding Maximally Large Homogeneous Segments

Here we present a criterion for selecting non-overlapping segments from the hierarchy (while in Sect. 8.3.2 we presented a way to select overlapping segments from the hierarchy). We start from the segmentation hierarchy $\{\mathcal{D}_i^h\}$ defined in Sect. 8.3.1. We define the heterogeneity of a segment $\mathcal{D}_i^h$ by the maximum difference of the statistics of its neighboring descendant nodes at level $h = 1$. More precisely, we define the heterogeneity of segment $\mathcal{D}_i^h$ to be:

$$\max_{j,k \in Des(\mathcal{D}_i^h)} ||\vec{S}_j^1 - \vec{S}_k^1||, \ \forall\ d_G(j,k) \leq 2, \tag{8.3}$$

where $d_G(j,k)$ is the graph distance between j,k at level $h = 1$ (i.e., we evaluate only the 1st and 2nd neighbors). This criterion considers homogeneous those segments whose statistics at level $h = 1$ change smoothly across the segment. This typically happens in large segments like sky, roads, animals. Heterogeneous segments will be those which have an abrupt change in their statistics.

We then fix a threshold t_{max} and generate an image partition

$$p_{t_{max}}(I(x)) \subset \{\mathcal{D}_i^h\}, \tag{8.4}$$

containing the biggest segments whose heterogeneity is less than t_{max}. This can be done by starting at the top-level $h = H$, keeping any node whose heterogeneity is less than t_{max}, proceeding to the child nodes otherwise, and continuing down the hierarchy until we reach levels where the heterogeneity threshold is achieved. Thus, the result is a set of non-overlapping segments covering the whole image space. Note that this is different from the entropy gain criterion used in Sect. 8.3.2, which allows to select overlapping segments, as interesting structures can happen at different levels (e.g., windows as a subpart of house).

8.3.3.2 Base Modeling and Detail

We assume that the image can be expressed as $I(x) = B(x) + R(x)$ where x is 2D position, $B(x)$ is base and $R(x)$ is detail (residual of the base). Both of them include all image channels. We assume that the base is spatially smooth within each maximally large homogeneous segment and, in particular, that its color intensity can be modeled by a low-order polynomial. We make no assumption about the spatial form of the detail. (Note that for intrinsic images it is typically assumed that the shadows are spatially smooth while the texture/albedo is more jagged.)

More precisely, we define the base color of a segment by a polynomial approximation $b_k(\vec{x}_i, \vec{\omega})$ of order k, where $k \leq 3$. See examples in Fig. 8.5. We apply the polynomial approximation on each channel separately. The number of parameters $\vec{\omega}$ depends on the order of the polynomial and we use model selection to decide the

Fig. 8.5 Examples of polynomial base approximations. *Left*: original. *Center*: 0-order approximation (i.e., mean). *Right*: 0-order to 3rd-order approximation

order for each segment (we must avoid fitting a high-order polynomial to a small segment). These polynomial approximations are of form:

$$
\begin{aligned}
& b_k(\vec{x},\vec{\omega}) = \vec{x}^T\vec{\omega}, \qquad (8.5)\\
& k=0: \ \vec{x}=1, \ \vec{\omega}=\omega_0\\
& k=1: \ \vec{x}=[1,x_1,x_2], \ \vec{\omega}=[\omega_0,\omega_1,\omega_2]\\
& k=2: \ \vec{x}=[1,x_1,x_2,x_1^2,x_2^2,x_1x_2], \ \vec{\omega}=[\omega_0,\cdots,\omega_5]\\
& k=3: \ \vec{x}=[1,x_1,x_2,x_1^2,x_2^2,x_1x_2,x_1^3x_2^3,x_1x_2^2,x_2x_1^2], \ \vec{\omega}=[\omega_0,\cdots,\omega_9]
\end{aligned}
$$

The estimation of the parameters $\vec{\omega}$ of the polynomial is performed by linear least squares QR factorization [24]. The order k is selected based on the error, with a regularization term biasing towards lower order. See Fig. 8.6-right. The regularization is weighted by ζ, whose value is not critical (it is set to produce models of all orders k, and not only $k=3$). In a given segment we have a set of pixels with 2D positions x and color intensity values $I_c(x)$. For a given channel c of the segment $\mathcal{D}_i^h$, we minimize:

$$
\min_{\vec{\omega},k} \sum_{x\in\mathcal{D}_i^h} \left(I_c(x) - b_k(\vec{x},\vec{\omega})\right)^2 + \zeta k. \qquad (8.6)
$$

Fig. 8.6 Different segments can have different polynomial order k. *Left*: original. *Center*: polynomial base approximation. *Right*: order of the polynomial, where: *dark-blue*: $k = 0$, *light-blue*: $k = 1$, *yellow*: $k = 2$; *red*: $k = 3$

Fig. 8.7 Example of detail (*right image*) in front of different appearance segments: sky, road, and building

We estimate the base $B_c(x)$ of each color channel c for the whole image by fitting the polynomial for each maximally large homogeneous segment. Then, we estimate the detail to be the residual $R_c(x) = I_c(x) - B_c(x)$.

Our current method works well in most cases, see Fig. 8.7, but it is not appropriate for segments where the amount of detail is similar to the amount of base appearance. This happens, for example, for an image of a leafy tree with blue sky behind it. Such situations require a more complex model which has a prior on the details and allows the base to be fit by a more flexible function (but still smooth).

"Base-detail" provides a unified model for several visual tasks that are often modeled separately. These include: (I) Elementary tasks such as gain control, which converts the large dynamic range of luminances into a smaller range of intensities which can be encoded by neurons and transmitted to the visual cortex. A standard hypothesis is that it is performed by ganglion cells in the retina, by Difference of Gaussian, or Laplacian of Gaussian [39], filters to preserve the contrast while removing the base. From our perspective, the contrast is the detail. (II) Decomposition of intensity into albedo and shading patterns as required by shape from shading algorithms [26, 27] when used to construct the 2 1/2 sketch [39] or intrinsic image [9]. The difference is that we do not estimate 3D geometry, noting that intrinsic image models make strong assumptions about images which are often

invalid (e.g., smooth intensity patterns can be due to light sources at finite distance and not to the geometry of the viewed surface). (III) Separation of texture from background. Here the detail represents the texture patterns, e.g., the blades of grass while the base is a smooth green intensity pattern. (IV) Decomposing images into frequency components. In this case, the detail is analogous to the high-frequencies. But frequency analysis is based on linear analysis of images while our approach is inherently nonlinear because it involves segmentation. (V) Image compression. Base details suggests a strategy where the base efficiently encodes the rough appearance and the detail encodes the rest. It captures the natural intuition that regions which have a lot of detail are harder to compress.

8.3.4 Image Enhancement

We illustrate how base – detail decomposition can be used for image enhancement. In Fig. 8.8-bottom, we plot $B(x) + \vartheta R(x)$, for different ϑ values, where ϑ is a parameter indicating the amount of enhancement. Another example is shown in Fig. 8.9. Our approach opens the doors to segment-wise manipulation, which is useful in common situations like when segments have different illumination.

Note that the widely used bilateral filter [44] is very local compared to our segmentation-based approach. In Fig. 8.10 we show an example of base-detail decomposition produced by bilateral filtering. For example, the top-right cloud in the image cannot be separated as a detail by the bilateral filter, but it is successfully separated as detail using our approach.

Fig. 8.8 *Top*: Original $I(x)$, base $B(x)$, detail $R(x)$, detail magnitude $||R(x)||_2$ for better visualization. *Bottom*: Base + detail $B(x) + \vartheta R(x)$, with different amounts of detail, $\vartheta = \{0.5, 1, 2, 4\}$

Fig. 8.9 Example of enhanced image. Weak details can be multiplied to become more visible with respect to the base

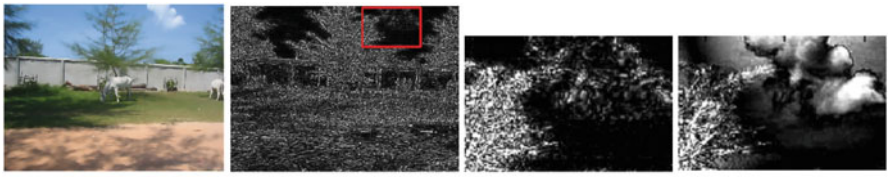

Fig. 8.10 Limitations of bilateral filter. From left to right: Bilaterally filtered (BLF) image; Residual (detail) of the bilateral filtering; Zoom-in of the residual; Zoom-in of the detail that our segmentwise base-detail decomposition produces

8.3.5 Saliency

Images of three-dimensional scenes contains structure at different scales and resolutions. Humans often need to foveate specific image locations to acquire higher level of details. For example, a small image blob might correspond to a person walking towards you and require further investigation. In this work we consider only bottom-up attention where fixation saliency is used to predict the first few seconds (3 s) of free-viewing of an image. The prediction consists of a probability map which does not take order of fixation into account. (By contrast, in top-down attention humans actively search for specific objects or scene structures.)

Our saliency model takes as input the base-detail decomposition $B_c(x), R_c(x)$, generated for a partition $p_{t_{max}}(I(x))$, defined in (8.4), whose minimum homogeneity threshold is t_{max} (Sect. 8.3.3). Note that candidate regions are not used here. Each image pixel x is assigned to the segment $i(x)$ which contains it and we define $|\mathcal{D}_{i(x)}| = size(\mathcal{D}_i)$, if $x \in \mathcal{D}_i$, where $\{\mathcal{D}_i \in p_{t_{max}}(I(x))\}$ are the segments of the partition. Similarly, we evaluate each segment's average detail and assign this value to all pixel positions of the segment support, obtaining $A(x) = \dfrac{1}{size(\mathcal{D}_i)} \displaystyle\sum_{z \in \mathcal{D}_i} R^A(z)$ if $x \in \mathcal{D}_i$. Here, $R^A(z)$ is the mean of the detail's $n_c = 3$ color channels at position

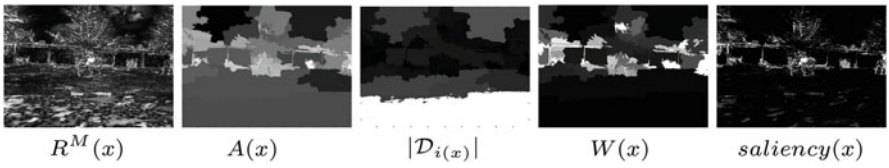

Fig. 8.11 From left to right: Maximum-channel detail $R^M(x)$; segmentwise average detail $A(x)$; segment size $|\mathcal{D}_{i(x)}|$; weight factor $W(x) = \sqrt{A(x)/|\mathcal{D}_{i(x)}|}$; $saliency(x) = R^M(x)[(1-\gamma)W(x)+\gamma]$

Fig. 8.12 An illustration of how our saliency model penalizes the detail in large roughly-homogeneous segments. The representation on the *right* is obtained by $I'_c(x) = B_c(x)+W(x)R_c(x)$, for each color channel c

z, that is, $R^A(z) = \frac{1}{n_c}\sum_{c=1}^{n_c} R_c(z)$. We use the segment sizes and the segmentwise average detail to weight the maximum-channel detail $R^M(x) = \max_{c=1}^{n_c} R_c(x)$ (see Fig. 8.11). The weight we propose is given by $W(x) = \sqrt{A(x)/|\mathcal{D}_{i(x)}|}$.

$$saliency(x) = R^M(x)\left[(1-\gamma)\,W(x)+\gamma\right]. \tag{8.7}$$

Here, γ is a small number, $\gamma = 0.15$ in our experiments. It allows to keep a fraction of $R^M(x)$ unweighted. This is useful for pixels whose weight is close to zero, $W(x) \approx 0$. The γ parameter means that the detail $R(x)$ is never completely ignored.

Intuitively, we relate the detail (Fig. 8.11-left) to bottom-up saliency. However, we penalize detail which belongs to large segments, without eliminating it completely (Fig. 8.11-right). An illustration of how an image would look like with this kind of detail penalization is shown in Fig. 8.12. The use of the segment size as an important saliency factor could be related to figure-ground pre-attentive mechanisms in V1. In terms of V1 neuron responses, very small regions tend to be highlighted against larger regions [50], but in this paper we do not address neurophysiology.

Our hypothesis is that regions which cannot be described by a simple model require foveation. This is the case of small regions with a lot of detail. The segments that are less likely to require foveation are those which are fit well by a simple

polynomial model (have little detail), as well as those which have detail but are large. In the latter case, the detail is likely to be due to a texture pattern, e.g., grass.

Classical models (e.g., [30]) use multiscale processing. Instead of this, our segment-based approach adapts to local scales of images. Also, unlike classical models, we do not explicitly use neural mechanisms such as center-surround receptive fields and lateral inhibition mechanisms. But it can be argued that base-detail decomposition is implicitly accomplishing similar functions.

The proposed fixation saliency method predicts human fixations. Note that this is different from salient object proposals. It is possible to link human fixations predictions and candidate regions by machine learning, as shown in [38]. But in this work we do not address this issue.

8.4 Experiments

In this section, we present results of the candidate region proposals (Sect. 8.3.2) and the bottom-up saliency (Sect. 8.3.5) as a prediction of free-viewing human fixations. Both of them are based on the bottom-up segmentation we propose (Sect. 8.3.1). The fundamental theory behind the saliency method is the base-detail decomposition (Sect. 8.3.3). We do not evaluate base-detail decomposition and image enhancement because there is no natural way of doing it. We do not focus on image segmentation, so we do not include experiments on it.

8.4.1 Datasets

Many of the classic datasets are biased because they were collected with a specific purpose, i.e., for saliency experiments. They are mostly composed of iconic photographs, presenting a clearly salient and centered object over a simple background. But this is highly atypical of natural images, which typically include many objects with complex relations and partial occlusions (humans rarely see iconic images). Hence it arguably more realistic to study saliency on natural image datasets such as Pascal, which has been one of the leading reference benchmark in Computer Vision for the last years. Recently, Hou et al. [38] released the free-viewing fixations of 8 participants on a subset of 850 images of Pascal (first 3 s). In this subset we have an average of 5.18 foreground objects per image and an average of 2.93 background objects. An extreme case is the rightmost image in Fig. 8.13, which has 52 foreground objects, most of which are far from the center of the image. A representative case is the third from the right image in Fig. 8.13, with 6 foreground objects.

For our candidate regions experiments, we use a subset of 1,288 images of Pascal VOC, for comparison with [6], as detailed in [10]. For the bottom-up saliency experiments, we use the 850 images of Pascal-S which include human fixations. We

Fig. 8.13 (**a**) Examples of iconic images from ImgSal [36]; (**b**) Examples from a non-iconic dataset: Pascal VOC [17]

also experiment on the 1,003 images of the standard dataset Judd [31], which can be considered non-iconic, although we have no statistics of the number of objects or their distribution in the images.

8.4.2 Candidate Regions

In this section, we evaluate the coverage of our candidate regions. Initially, we obtained an average of 116 selected-segments per image after selection and from these we make an average of 721 combinations which constitute the pool of candidate regions per each image. The evaluation metric is Intersection over Union (IoU), which accounts the number of pixels of the intersection between a candidate region and a groundtruth region, divided by the number of pixels of their union.

We evaluate the generated candidate regions with the 57-classes ground truth, containing both foreground and "stuff" classes. We compare our Candidate Regions (CR) to three state-of-the-art methods. (I) The classical Constrained Parametric Min-Cuts (CPMC) [15] method is designed for foreground objects, which explains its better performance on foreground objects. Their overall performance on the 57 classes is lower than our performance. (II) In [6], the segment combinations are generated by taking combinations of the 150 segments (on average) that their hierarchical segmentation approach outputs for each image. Their method is more sophisticated than ours and we observe that they tend to get larger and less homogeneous segments than we do. Our performance is lower but comparable: 74 % IoU versus 77 % for [6]. But we achieve it with nearly half the number of combinations – 721 compared to 1,322 – and with a simpler and faster algorithm (4 s per image in its Matlab prototype). In Table 8.1 we refer to their segments as UCM-combs and to our candidate regions as CR-combs. (III) The Selective Search [46] method is competitive in terms of speed. Our method outperforms theirs on the region candidates task (74.0 % compared to 67.8 % IoU), with less than half the number of proposals. (Note, however, that [46] present results for bounding boxes and not for regions). See an example of the proposals generated by our CR-combs method in Fig. 8.14. See Table 8.1 with the region-based IoU and recall results.

Table 8.1 Region-based IoU (in %) comparison. CPMC [15], UCM [6], Sel. Search [46], and our CR – candidate regions. Boldface denotes the first and second best results

	All IoU	Recall (%)	# cands.	Time (s)
CPMC	59.6	57.6	**150**	250
UCM-combs	**77.0**	**80.0**	1,322	850
Sel. search	67.8	66.1	2,100	**4**
Our CR combs	**74.0**	**70.3**	**721**	**4**

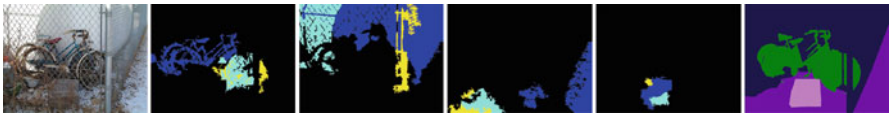

Fig. 8.14 Left-to-right and top-to-bottom: Original image, top three segments for bike, wall, snow, rock, and ground truth. Note that the segments are good even for object classes that perform poorly overall (e.g., bike)

8.4.3 Saliency

The fixation saliency method that arises from our unified approach predicts free-viewing human fixations surprisingly well. Despite only accounting for saliency within segments and not taking into account inter-segment saliency, our method is among the highest ones in complex datasets like Pascal-S [38] and Judd [31]. Pascal is a particularly interesting case because state-of-the-art fixations methods have low performance on it (perhaps because they were developed and tested for iconic images). In Pascal our method outperforms the state of the art. On the Judd dataset only AWS [21] outperforms our method.

In Fig. 8.15 we show a comparison of our Base-Detail Saliency (BDS) method, Adaptive Whitening Saliency, AWS [21], Image Signature, SIG [28], and L. Itti's original model [30]. In Fig. 8.16 we show some examples for qualitative comparison between the results of the different algorithms.

It is hard to determine the failure modes of our saliency algorithm. Our reliance on segmentation may seem problematic. It is known that segmentation is an ill-posed problem and no low-level segmentation algorithm exists that can reliably detect the boundaries of objects without top-down assistance. But our approach is more robust because we rely only on a proto-segmentation. Still errors in the segmentation can cause errors in the base-detail decomposition which may cause our approach to fail.

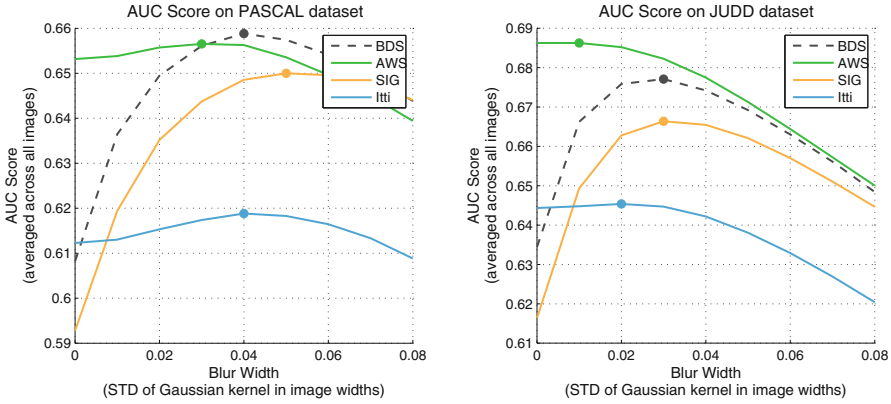

Fig. 8.15 Bottom-up saliency performance. *Left*: Pascal-S dataset [38]. *Right*: Judd dataset [31]. Approaches compared: Our Base-Detail Saliency (BDS), Adaptive Whitening Saliency (AWS, [21]), Image Signature (SIG, [28]), L. Itti's original model (Itti, [30])

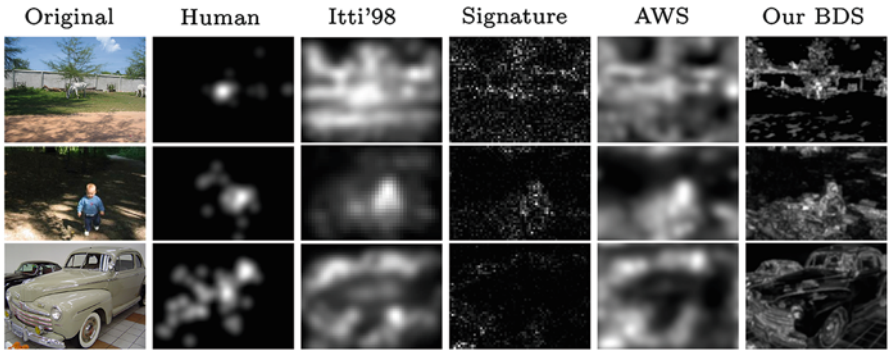

Fig. 8.16 From left ro tight: Original; Human – fixations collected on 8 subjects with free-viewing task, first 3 s [38]; Itti's original model [30]; Spectral signature [28]; AWS [21]; Our Base-Detail Saliency (BDS) Bonev and Yuille

8.5 Conclusions

We propose a unified approach addressing a set of early-vision bottom-up processes: segmentation, candidate regions, base-detail decomposition, image enhancement, and saliency for fixations prediction.

Our unified approach allows the segmentwise decomposition of the image into "base" and "detail". This proves to be more versatile than a local smoothing of the image. It provides directly for image enhancement, for a novel model of fixation saliency. It is related to other vision topics which are usually formalized as different problems.

We show state-of-the-art results on our candidate regions and on our saliency for free-viewing fixation prediction. For the latter we use the psychophysics data available for the Pascal VOC dataset, which is non-iconic and particularly difficult for the state-of-the-art saliency algorithms.

Acknowledgements We would like to thank Laurent Itti, Li Zhaoping, John Flynn, and the reviewers for their valuable comments. This work is partially supported by NSF award CCF-1317376, by ONR N00014-12-1-0883 and by NVidia Corp.

References

1. Achanta R, Shaji A, Smith K, Lucchi A, Fua P, Susstrunk S (2012) SLIC superpixels compared to state-of-the-art superpixel methods. TPAMI 34(11):2274–2282
2. Alexe B, Deselaers T, Ferrari V (2012) Measuring the objectness of image windows. TPAMI 34(11):2189–2202
3. Alpert S, Galun M, Brandt A, Basri R (2012) Image segmentation by probabilistic bottom-up aggregation and cue integration. TPAMI 34(2):315–327
4. Arbelaez P (2006) Boundary extraction in natural images using ultrametric contour maps. In: Proceedings of the 2006 conference on computer vision and pattern recognition workshop, CVPRW '06. IEEE Computer Society, Washington, DC, pp 182–
5. Arbelaez P, Maire M, Fowlkes C, Malik J (2011) Contour detection and hierarchical image segmentation. TPAMI 33(5):898–916
6. Arbelaez P, Hariharan B, Gu C, Gupta S, Malik J (2012) Semantic segmentation using regions and parts. In: CVPR, Providence
7. Bae S, Paris S, Durand F (2006) Two-scale tone management for photographic look. ACM Trans Graph 25(3):637–645
8. Barron JT, Malik J (2012) Color constancy, intrinsic images, and shape estimation. In: ECCV, Florence
9. Barrow HG, Tenenbaum JM (1978) Recovering intrinsic scene characteristics from images. Technical report 157, AI Center, SRI International, 333 Ravenswood Ave., Menlo Park, CA 94025
10. Bonev B, Yuille AL (2014) A fast and simple algorithm for producing candidate regions. In: European conference on computer vision (ECCV 2014), Zurich
11. Borji A, Itti L (2013) State-of-the-art in visual attention modeling. IEEE Trans Pattern Anal Mach Intell 35(1):185–207
12. Borji A, Sihite DN, Itti L (2013) Objects do not predict fixations better than early saliency: a re-analysis of Einhäuser et al.'s data. J Vis 13(10):18
13. Borji A, Cheng M, Jiang H, Li J (2014) Salient object detection: a survey. CoRR, abs/1411.5878
14. Bradley C, Abrams J, Geisler WS (2014) Retina-v1 model of detectability across the visual field. J Vis 14(12):22
15. Carreira J, Sminchisescu C (2012) CPMC: automatic object segmentation using constrained parametric min-cuts. TPAMI 34(7):1312–1328
16. Einhäuser W, Spain M, Perona P (2008) Objects predict fixations better than early saliency. J Vis 8(14):18
17. Everingham M, Van Gool L, Williams CK, Winn J, Zisserman A (2010) The PASCAL visual object classes (VOC) challenge. Int J Comput Vis 88(2):303–338
18. Farbman Z, Fattal R, Lischinski D, Szeliski R (2008) Edge-preserving decompositions for multi-scale tone and detail manipulation. ACM Trans Graph 27(3):67:1–67:10

19. Felzenszwalb PF, Huttenlocher DP (2004) Efficient graph-based image segmentation. IJCV 59(2):167–181
20. Galun M, Sharon E, Basri R, Brandt A (2003) Texture segmentation by multiscale aggregation of filter responses and shape elements. In: ICCV '03, Nice, pp 716–
21. Garcia-Diaz A, Leborán V, Fdez-Vidal XR, Pardo XM (2012) On the relationship between optical variability, visual saliency, and eye fixations: a computational approach. J Vis 12(6):1–22
22. Geman S, Geman D (1984) Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. IEEE Trans Pattern Anal Mach Intell 6(6):721–741
23. Gollisch T, Meister M (2010) Eye smarter than scientists believed: neural computations in circuits of the retina. Neuron 65(2):150–164
24. Golub GH, Van Loan CF (2012) Matrix computations, vol 3. JHU Press, Baltimore
25. Gonzalez RC, Woods RE, Eddins SL (2004) Digital image processing using matlab. Pearson Prentice Hall, Upper Saddle River
26. Gorelick L, Basri R (2009) Shape based detection and top-down delineation using image segments. Int J Comput Vis 83(3):211–232
27. Horn BKP, Brooks MJ (1986) The variational approach to shape from shading. Comput Vis Graph Image Process 33(2):174–208
28. Hou X, Harel J, Koch C (2012) Image signature: highlighting sparse salient regions. IEEE TPAMI 34(1):194–201
29. Humayun A, Li F, Rehg JM (2014) RIGOR: reusing inference in graph cuts for generating object regions. In: Proceedings of IEEE conference on computer vision and pattern recognition (CVPR), Columbus. IEEE, New York
30. Itti L, Koch C, Niebur E (1998) A model of saliency-based visual attention for rapid scene analysis. IEEE TPAMI 20(11):1254–1259
31. Judd T, Ehinger K, Durand F, Torralba A (2009) Learning to predict where humans look. In: ICCV, Kyoto, pp 2106–2113. IEEE, New York
32. Land EH (1977) The retinex theory of color vision. Sci Am 237(6):108–28
33. Leclerc YG (1989) Image and boundary segmentation via minimal-length encoding on the connection machine. In: Proceedings of a workshop on image understanding workshop, Palo Alto. Morgan Kaufmann, San Francisco, pp 1056–1069. ISBN 1-55860-070-1. http://dl.acm.org/citation.cfm?id=94703.99744
34. LeCun Y, Bottou L, Bengio Y, Haffner P (1998) Gradient-based learning applied to document recognition. Proc IEEE 86:2278–2324
35. Leonenko N, Pronzato L, Savani V (2008) A class of Rényi information estimators for multidimensional densities. Ann Statist 36(5):2153–2182
36. Li J, Levine M, An X, He H (2011) Saliency detection based on frequency and spatial domain analyses. In: Proceedings of BMVC, Dundee, pp 86.1–86.11. http://dx.doi.org/10.5244/C.25.86
37. Li J, Levine MD, An X, Xu X, He H (2013) Visual saliency based on scale-space analysis in the frequency domain. IEEE Trans Pattern Anal Mach Intell 35(4):996–1010
38. Li Y, Hou X, Koch C, Rehg JM, Yuille AL (2014) The secrets of salient object segmentation. In: CVPR, Columbus
39. Marr D (1982) Vision: a computational investigation into the human representation and processing of visual information. Henry Holt and Co., New York
40. Mottaghi R, Chen X, Liu X, Fidler S, Urtasun R, Yuille A (2014) The role of context for object detection and semantic segmentation in the wild. In: CVPR, Columbus
41. Russ JC, Woods RP (1995) The image processing handbook. J Comput Assist Tomogr 19(6):979–981
42. Shapley R, Enroth-Cugell C (1984) Visual adaptation and retinal gain controls. Prog Retin Res 3:263–346
43. Todorovic S, Ahuja N (2008) Region-based hierarchical image matching. IJCV 78(1):47–66
44. Tomasi C, Manduchi R (1998) Bilateral filtering for gray and color images. In: Sixth international conference on computer vision, 1998. IEEE, Washington, DC, pp 839–846

45. Tu Z, Zhu S-C, Shum H-Y (2001) Image segmentation by data driven Markov chain Monte Carlo. In: Proceedings of eighth IEEE international conference on computer vision, 2001. ICCV 2001, Vancouver, vol 2, pp 131–138
46. Uijlings JRR, van de Sande KEA, Gevers T, Smeulders AWM (2013) Selective search for object recognition. Int J Comput Vis 104(2):154–171
47. Woodham RJ (1980) Photometric method for determining surface orientation from multiple images. Opt Eng 19(1):191139–191139
48. Xu C, Xiong C, Corso JJ (2012) Streaming hierarchical video segmentation. In: ECCV, Florence
49. Yuan L, Sun J (2012) Automatic exposure correction of consumer photographs. In: Fitzgibbon AW, Lazebnik S, Perona P, Sato Y, Schmid C (eds) ECCV (4). Volume 7575 of Lecture notes in computer science. Springer, Berlin/New York, pp 771–785
50. Zhaoping L (2003) V1 mechanisms and some figure-ground and border effects. J Physiol 97(1):503–515
51. Zhaoping L (2014) Understanding vision: theory, models, and data. Oxford University Press, Oxford
52. Zhu SC, Yuille A (1996) Region competition: unifying snakes, region growing, and Bayes/MDL for multiband image segmentation. IEEE Trans Pattern Anal Mach Intell 18(9):884–900
53. Zhu L, Chen Y, Lin Y, Lin C, Yuille A (2012) Recursive segmentation and recognition templates for image parsing. IEEE Trans Pattern Anal Mach Intell 34(2):359–371
54. Zhu Y, Zhang Y, Yuille A (2014) Single image super-resolution using deformable patches. In: 2014 IEEE conference on computer vision and pattern recognition (CVPR), Columbus, pp 2917–2924
55. Zitnick CL, Dollár P (2014) Edge boxes: locating object proposals from edges. In: ECCV, Zurich