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[†]

University of California, Los Angeles

March 17, 2015

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 segment-wise analysis. Further, we argue that some visual tasks require the ability to decompose the appearance of the segments into "base" appearance and "detail" appearance. An important, and surprising, result of these decomposition is the ability to successfully predict 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, to obtain a higher resolution representation. 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.

1 Introduction

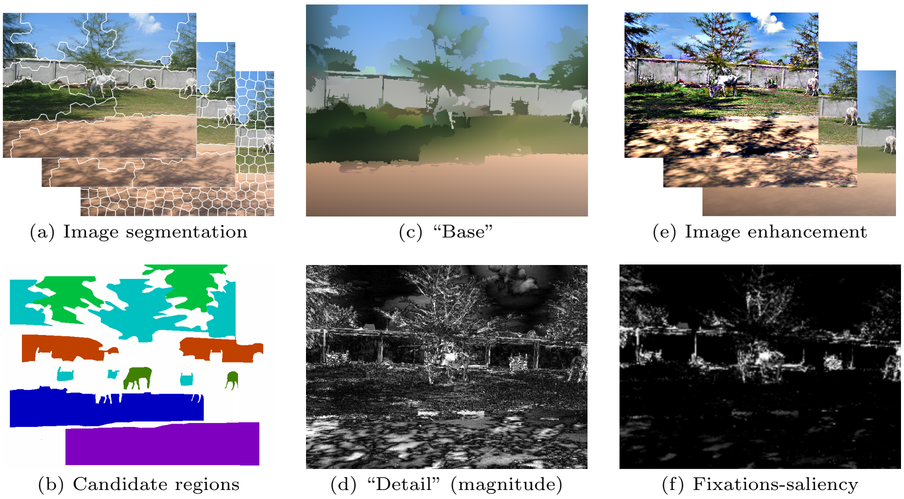
Low level vision is visual processing that treats images as patterns and makes no specific assumptions about the objects that might be present or the structure of the scene. In short, the processing is generic and intended to be suitable for all images, regardless of their semantic content or high level layout. Examples of low-level vision tasks include segmentation, candidate regions or object proposals, image enhancement. Low-level processing is typically performed in preparation for high-level tasks, and it can be used to allocate computational

[*] E-mail: bonev@ucla.edu; Corresponding author

[†] E-mail: yuille@stat.ucla.edu

resources for more detailed processing. It might be thought of as analogous to the processing in the retina and the V1 visual cortex of the brain.

In this paper we propose a unified framework for several low-level vision tasks (Fig. 1) that are typically modeled separately. These tasks include the generation of segments at multiple levels, candidate regions for object recognition, base-detail decomposition – where an image is decomposed into a visual summary plus fine details, image enhancement and the prediction of human fixations.



(a) Image segmentation

(c) "Base"

(e) Image enhancement

(b) Candidate regions

(d) "Detail" (magnitude)

(f) Fixations-saliency

Figure 1: We propose a unified approach for a set of bottom-up vision processes: (a) image segmentation – a hierarchy of image partitions at multiple levels; (b) candidate regions – a pool of possibly overlapping proposals for further evaluation by object recognition methods (best candidates illustrated); (c,d) "base-detail" decomposition – modeling of the appearance and its residual, 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.

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 larger and less homogeneous. 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" (Bonev and Yuille, 2014). We refer to them as candidate regions. They consist of a pool of 500 to 1500 regions which are further evaluated by a higher-level method (out of the scope of this paper). The high-level method assigns them category-specific scores in order 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 all the texture and details. The image is the sum of both. More precisely, the base is obtained by fitting simple appearance models (polynomials) to the image segments. The detail is the residual. See an example in Fig. 1-(c,d) and in Fig. 8. This base-detail decomposition enables us to process the image in several ways, such as enhancing the details and/or the base, or removing the shadows (details) from a grass lawn (base). In addition, we can compute saliency for predicting free-viewing human fixations.

It is well known that when humans examine an image they do not gaze on it uniformly, instead they fixate on certain parts of the image. The fixation saliency model we propose prioritizes small segments which have strong detail. 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, while in small segments the detail may correspond to structures which require more detailed models to process. Our fixation saliency model predicts human fixations with a state-of-the-art performance on complex datasets, like Pascal (Everingham et al., 2010) and Judd (Judd et al., 2009), for which we present evaluations.

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, and do not explore biological evidences supporting our work. We are focusing on complex visual scenes, instead of artificial stimuli, because we think it is important to model visual abilities in real-world conditions.

2 Background and related literature

There is an enormous literature on segmentation. Markov Random Fields, where the appearance of the image is encoded in unary and pairwise terms, have long been used (Geman and Geman, 1984), as well as hierarchical segmentations (Zhu et al., 2012) by recursive segmentation, oversegmentation methods (Achanta et al., 2012) by local k-means clustering, and graph-based methods (Felzenszwalb and Huttenlocher, 2004). In this paper, we address hierarchies and so the most relevant work is Segmentation by Weighted Aggregation, (SWA), (Alpert et al., 2012) and Ultrametric Contour Map, (UCM), (Arbelaez, 2006). Related hierarchical grouping methods include the SWA algorithm (Galun et al., 2003) and a recent variant (Alpert et al., 2012). This method has also been extended to video segmentation (Xu et al., 2012).

Candidate regions and salient objects proposals are increasingly topical in computer vision because it makes proposals for the likely positions of objects which can then be analyzed in detail using more complex models, e.g., deep convolutional neural networks (LeCun et al., 1998). Our candidate regions approach differs from the methods in the literature in the goals: we propose regions for both objects and background regions or “stuff” (e.g., sky). The related work is recent. SWA (Galun et al., 2003) uses a local segment-saliency measure but it does not take the further step of grouping the segments into combinations (as done in (Arbelaez et al., 2011; Arbelaez et al., 2012)). We should mention hierarchical segmentation which has been used to learn models of objects (Todorovic and Ahuja, 2008). Most methods in the literature have been evaluated for finding segments which cover foreground objects (Alpert et al., 2012; Uijlings et al., 2013), while ours is for background classes as well. Finally, there are methods which differ in that they mainly exploit the edges instead of the appearance statistics (Carreira and Sminchisescu, 2012; Humayun et al., 2014; Zitnick and Dollár, 2014).

There is little work which directly addresses "base-detail" segmentation, but there is a large literature on related topics. In the digital image processing community there is a related concept, "base-detail", but it is extracted locally (Bae et al., 2006), by applying filters like bilateral filter. A related topic is gain control. It is plausible the primate visual systems also use this type of processing, perhaps in the retina, as part of sophisticated gain control (Shapley and Enroth-Cugell, 1984; Bradley et al., 2014). Enhancement of detail is also at the heart of many super-resolution methods (Zhu et al., 2014). There has been considerable work, in the shape from shading community, in decomposing intensity patterns into shading components and texture/albedo components (Barrow and Tenenbaum, 1978; Horn and Brooks, 1986; Gorelick and Basri, 2009). If multiple images are present of the same object under different lighting conditions, then this decomposition can be done by photometric stereo (Woodham, 1980). Otherwise, methods make prior assumptions that the shading component is spatially smooth while the albedo/texture pattern is more jagged (e.g., (Barron and Malik, 2012)). Similar assumptions are also applied to the classic Mondrian problem (Land, 1977).

Base-detail is closely related to image enhancement. In this field, segment-wise methods are not typically used (Russ and Woods, 1995; Gonzalez et al., 2004). Local filters are usually applied, like the bilateral filter in (Bae et al., 2006) or the weighted least squares (Farbman et al., 2008). There are some exceptions, like (Yuan and Sun, 2012) where segment-wise exposure correction is proposed.

By fixation saliency we refer to bottom-up saliency for human fixations prediction. It does not consider top-down processes involving cognitive factors, e.g., counting the animals in the image. For this reason, the psychophysics evaluations are performed for a free-viewing task. One of the first successful approaches was Itti's original model (Itti et al., 1998). An arguably simple method which produces state-of-the-art results is Image Signature (Hou et al., 2012). The best performing method we have found in the state of the art is the Adaptive Whitening Saliency (Garcia-Diaz et al., 2012). See the most recent review for more details (Borji and Itti, 2013). Finally, there are works (Li et al., 2013; Li et al., 2014) linking human vision saliency with candidate regions to predict salient objects (Borji et al., 2014). Salient objects differ from fixation saliency in that the whole object is accounted as salient based on the fixations that fall within it, while in fixation saliency only the coordinates are accounted for, producing a fixations map (in section 4.3, see Fig. 16, second column). The fixation saliency models we propose is based on base-detail decomposition, which makes it substantially different from the literature. It performs at the level of the best methods in the state of the art.

Finally, recent biological vision studies suggest that early visual processing is more sophisticated than traditional models of the retina and V1, which were built mainly on spatiotemporal filters. For example, studies of the retina suggest that it is "smarter than scientists believed" (Gollisch and Meister, 2010). It is also plausible that tasks such as fixation saliency are computed in V1 (Zhaoping, 2014).

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 define them in terms of a unified approach of bottom-up vision processing.

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 $\{\mathcal{D}_i : i = 1, ..., n\}$ 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 $h = 1, ..., H$. 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 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. 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 (Arbelaez, 2006). Fig. 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. 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 (Leclerc, 1989; Zhu and Yuille, 1996; Tu et al., 2001) which propose minimizing a cost function of form:

$$E(\{\mathcal{D}_i\}, \{\vec{S}_i\}) = \sum_i \sum_{x \in \mathcal{D}_i} |\vec{S}_i - \vec{S}(x)|^2 - \lambda \sum_i \sum_{x \in \partial D_i} e(x). \qquad (1)$$

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 $\mathcal{D}_i$.

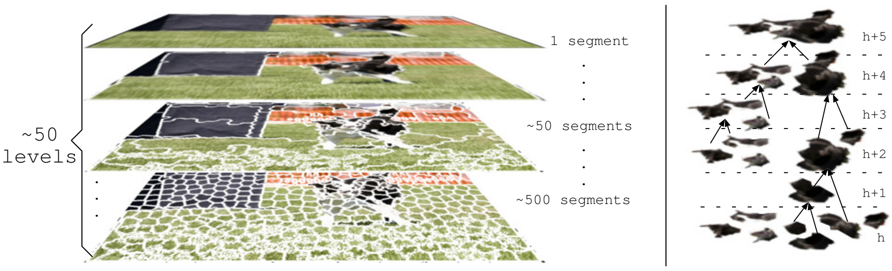


Figure 2: Left: Multiple levels in a hierarchy. Segments with a good coverage of objects or parts may happen at different levels. 80% to 90% of the segments can be discarded because they go across boundaries of objects or because they don't cover a large area of an object. Right: Segments at level $h+1$ are composed of one or two segments in level h.

We initialize our algorithm by using the SLIC (Achanta et al., 2012) algorithm to compute the lowest level, $h = 1$, of our hierarchy. Essentially, SLIC performs an expectation-minimization of (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 (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 (Bonev and Yuille, 2014).

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.

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* (Li et al., 2013; Alexe et al., 2012), where psychophysical studies show that humans have tendencies to look at salient objects (Einhäuser et al., 2008). Note that salient objects, however, do not predict human eye fixations well (Borji et al., 2013) and these can be better described by bottom-up saliency cues (Itti et al., 1998) 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 (Li et al., 2014).

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. 4. This intuition was validated (Bonev and Yuille, 2014) using the extended labeling of Pascal VOC (Mottaghi et al., 2014) 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:

$$g = \mathcal{H}(\mathcal{D}_m^{h+1}) - \left\{\mathcal{H}(\mathcal{D}_i^h) + \mathcal{H}(\mathcal{D}_j^h)\right\}. \tag{2}$$

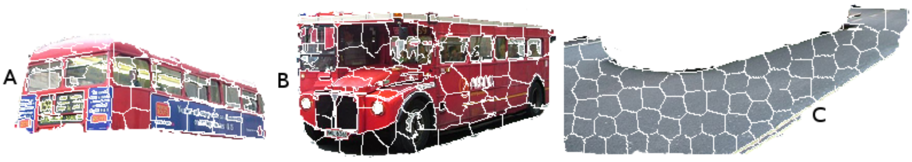


Figure 3: *Entropy gain* (section 3.2): When segments A and B are merged, the increase of entropy is not as big as if they were merged with C. *Homogeneity criterion* (section 3.3.1): Segment C is homogeneous. It presents smooth variation due to shading and lighting. Segments A and B are not homogeneous. Both entropy and homogeneity are calculated from the small (first level $\mathcal{D}_i^1$) segments, illustrated with white contours.

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. 3). The entropy is computed in a non-parametric manner (Bonev and Yuille, 2014) using the approximation proposed in (Leonenko et al., 2008). See an example of triplets of selected-segments in Fig. 4.

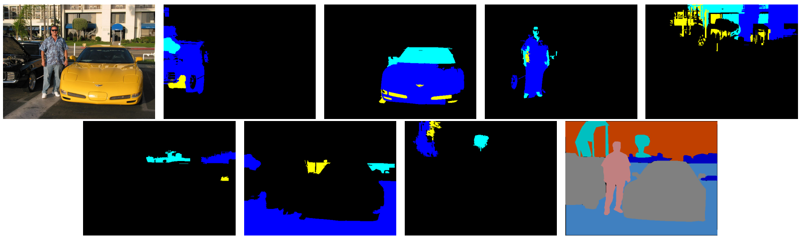


Figure 4: Examples of candidate regions for foreground and background regions. Left-to-right and top-to-bottom: image, top three selected-segments for left car, right car, person, building, grass, ground, trees, and ground truth. Most objects are covered well by two to three selected-segments.

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 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. However, in intrinsic images geometry and lighting sources are assumed. We do not consider such high-level concepts in our $I(x) = B(x) + R(x)$ decomposition (only in some very special situations the base and the detail of a segment may correspond to the shading and the albedo of an object). Thirdly, it also relates to transparency – e.g., the viewing of images through a dirty window – or when there is partial occlusion like tree leaves in front of a building. 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 (Bae et al., 2006) present in the image processing literature, which is obtained by local smoothing methods and not in a piecewise way.

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 and compose the base of these approximations (see section 3.3.2). We obtain the detail by computing the residual between the image and the interpolated color.

3.3.1 Finding maximally large homogeneous segments

Here we present a criterion for selecting non-overlapping segments from the hierarchy (while in section 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 section 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:

$$\max_{j,k \in Des(\mathcal{D}_i^h)} ||\vec{S}_j^1 - \vec{S}_k^1||, \ \forall \ d_G(j,k) \leq 2, \tag{3}$$

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

$$p_{t_{max}}(I(x)) \subset \{\mathcal{D}_i^h\}, \tag{4}$$

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 section 3.2, which allows to select overlapping segments, as interesting structures can happen at different levels (e.g., windows as a subpart of house).

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

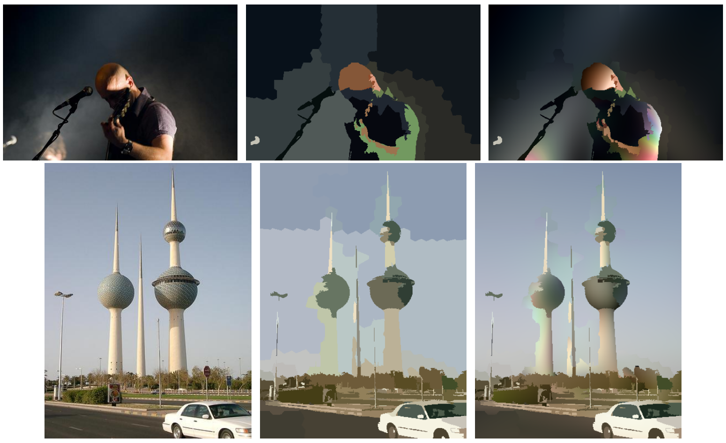


Figure 5: Examples of polynomial base approximations. Left: original. Center: 0-order approximation (i.e., mean). Right: 0-order to 3rd-order approximation.

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

$$\begin{aligned} b_k(\vec{x}, \vec{\omega}) \quad &= \vec{x}^T \vec{\omega}, && (5) \\ k = 0: \quad & \vec{x} = 1,\ \vec{\omega} = \omega_0 \\ k = 1: \quad & \vec{x} = [1, x_1, x_2],\ \vec{\omega} = [\omega_0, \omega_1, \omega_2] \\ k = 2: \quad & \vec{x} = [1, x_1, x_2, x_1^2, x_2^2, x_1 x_2],\ \vec{\omega} = [\omega_0, \cdots, \omega_5] \\ k = 3: \quad & \vec{x} = [1, x_1, x_2, x_1^2, x_2^2, x_1 x_2, x_1^3 x_2^3, x_1 x_2^2, x_2 x_1^2],\ \vec{\omega} = [\omega_0, \cdots, \omega_9] \end{aligned}$$

The estimation of the parameters $\vec{\omega}$ of the polynomial is performed by linear least squares QR factorization (Golub and Van Loan, 2012). The order k is selected based on the error, with a regularization term biasing towards lower order. See Fig. 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:

$$\min_{\vec{\omega}, k} \sum_{x \in \mathcal{D}_i^h} \left(I_c(x) - b_k(\vec{x}, \vec{\omega})\right)^2 + \zeta k. \qquad (6)$$

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 $R_c(x) = I_c(x) - B_c(x)$.

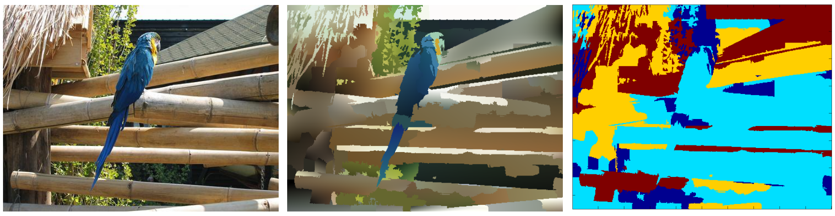


Figure 6: Different segments can have different polynomial order k. Left: original. Center: polynomial base approximation. Right: order of the polynomial, where: dark-blue: $k = 0$, light-blue: $k = 1$, yellow: $k = 2$; red: $k = 3$.

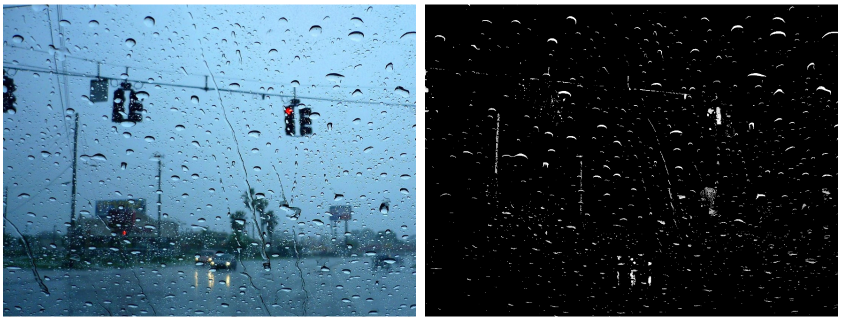


Figure 7: Example of detail (right image) in front of different appearance segments: sky, road, and building.

Our current method works well in most cases, see Fig. (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 (Marr, 1982), 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 (Horn and Brooks, 1986; Gorelick and Basri, 2009) when used to construct the 2 1/2 sketch (Marr, 1982) or intrinsic image (Barrow and Tenenbaum, 1978). The difference is that we do not estimate 3D geometry, noting that many intensity patterns are not due to shading. (III) The detail represents the texture patterns, e.g., details of grass while the base is a smooth intensity pattern of green. (IV) In terms of frequencies, the detail is the analogue to the high-frequencies. However, in "base-detail" segmentation is inherent to the decomposition and it has an impact on how coarse or fine the detail is in different segments. (V) In terms of information, the amount of detail of a segment is related to how much it can be compressed.

3.4 Image enhancement

We illustrate how base–detail decomposition can be used for image enhancement. In Fig. 8-bottom, we show the results of $B(x) + \vartheta R(x)$ for different ϑ values. See another example in Fig. 9. Our bottom-up approach opens the doors to segment-wise manipulation, which is necessary in common cases like having segments with different illumination.

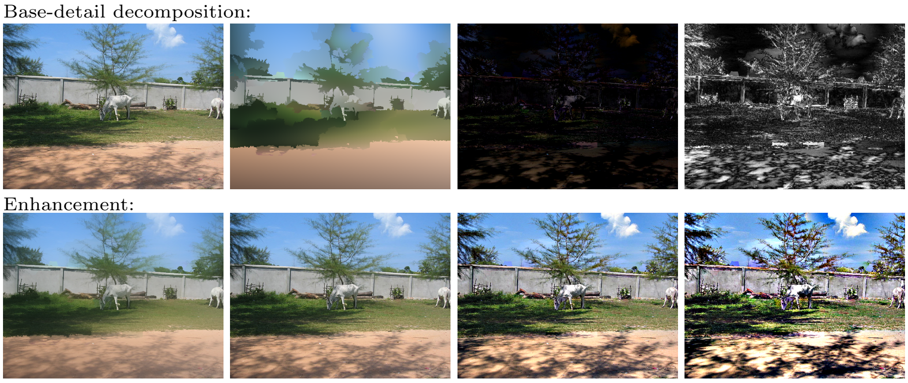


Figure 8: Top: Original $I(x)$, base $B(x)$, detail $R(x)$, detail magnitude $||R(x)||_2$ for better visualization. Bottom: Base + detail $B(x) + \vartheta R(x)$, with different amounts of detail, $\vartheta = \{0.5, 1, 2, 4\}$.

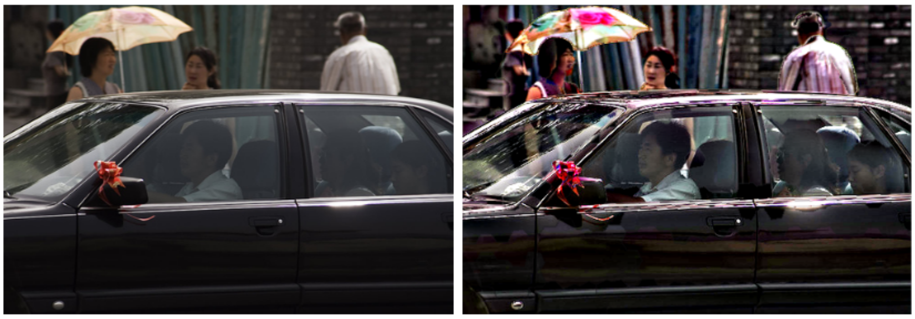
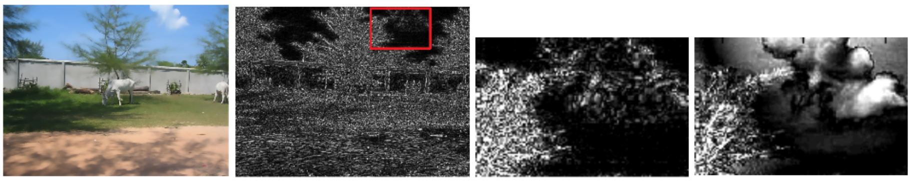


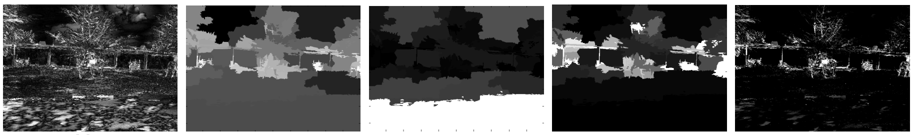
Figure 9: Example of enhanced image. Weak details can be multiplied to become more visible with respect to the base.

Note that the widely used bilateral filter (Tomasi and Manduchi, 1998) is too local compared to our segmentation-based approach. In Fig. 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 following our approach.



BLF's base BLF's detail BLF's detail zoom-in Our detail zoom-in

Figure 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.



$R^M(x)$ $A(x)$ $|\mathcal{D}_{i(x)}|$ $W(x)$ $saliency(x)$

Figure 11: From left to right: Maximum-channel detail $R^M(x)$; segmentwise average detail $A(x)$; segment size $|\mathcal{D}_{i(x)}|$; weight factor $W(x) = \sqrt{A(x)/|\mathcal{D}_{i(x)}|}$; $saliency(x) = R^M(x)[(1-\gamma)W(x) + \gamma]$.

3.5 Saliency

Images are perceived by our retina with a varying spatial resolution. Humans need to foveate specific image locations to acquire a higher level of detail. These locations highly depend on the task demands and on cognitive factors, but here we only consider bottom-up visual attention. This means that our fixation saliency is used to predict the first few seconds (3s) of free-viewing of an image. The prediction consists of a probability map, not considering order.

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 (4), whose minimum homogeneity threshold is t_{max} (section 3.3). Note that candidate regions are not used here. We assign to each image pixel x, the area size of the corresponding segment at that position, that is, $|\mathcal{D}_{i(x)}| = size(\mathcal{D}_i)$, 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 $A(x) = \dfrac{1}{size(\mathcal{D}_i)} \sum_{z \in \mathcal{D}_i} R^A(z)$ if $x \in \mathcal{D}_i$. Here, $R^A(z)$ is the mean of the detail's $n_c = 3$ color channels at position 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. 11). The weight we propose is given by $W(x) = \sqrt{A(x)/|\mathcal{D}_{i(x)}|}$.

$$saliency(x) = R^M(x)\left[(1-\gamma)\,W(x) + \gamma\right]. \qquad (7)$$

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 there is no case in which the detail $R(x)$ is to be completely ignored.

Intuitively, we relate the detail (Fig. 11-left) to bottom-up saliency. How-

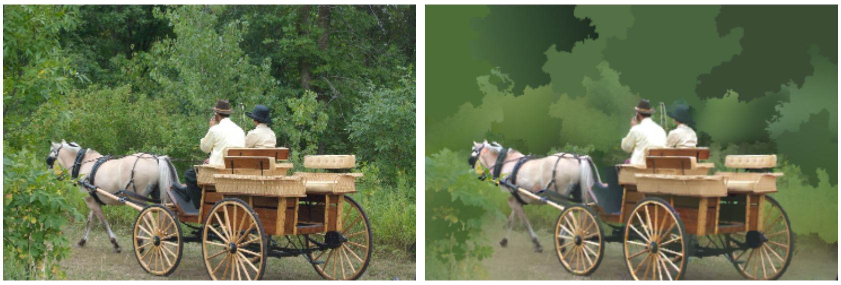


Figure 12: An illustration of how our saliency model penalizes the detail in large roughly-homogeneous segments. The representation on the right is obtained by $I'_c(x) = B_c(x) + W(x)R_c(x)$, on each color channel c.

ever, we penalize detail which belongs to large segments, without eliminating it completely (Fig. 11-right). An illustration of how an image would look like with this kind of detail penalization is shown in Fig. 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 (Zhaoping, 2003), 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., (Itti et al., 1998)) benefit from multiscale processing. Instead of this, our method makes use of segments of different sizes. Also, we do not explicitly model center-surround difference and cortical lateral inhibition mechanisms. It could be argued that the 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 (Li et al., 2014). In this work we don't address this question.

4 Experiments

In this section, we present results of the candidate region proposals (Subsection 3.2) and the bottom-up saliency (Subsection 3.5) as a prediction of free-viewing human fixations. Both of them are based on the bottom-up segmentation we propose (Subsection 3.1). The fundamental theory behind the saliency method is the base-detail decomposition (Subsection 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.

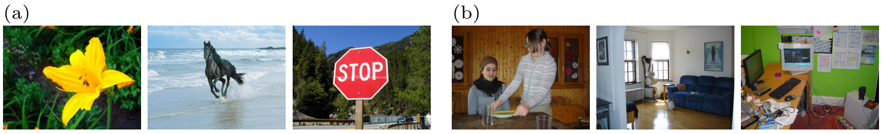


Figure 13: (a) Examples of iconic images from ImgSal (Jian Li and He, 2011); (b) Examples from a non-iconic dataset: Pascal VOC (Everingham et al., 2010).

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. However, the input of the human visual system consists of non-iconic images most of the time. Pascal is arguably a non-iconic dataset, and it has been the reference benchmark in Computer Vision for the last years. Recently, Hou et al. (Li et al., 2014) released the free-viewing fixations of 8 participants on a subset of 850 images of Pascal (first 3 seconds). 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. 13, which has 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. 13, with 6 foreground objects.

For our candidate regions experiments, we use a subset of 1,288 images of Pascal VOC, for comparison with (Arbelaez et al., 2012), as detailed in (Bonev and Yuille, 2014). For the bottom-up saliency experiments, we use the 850 images of Pascal-S which include human fixations. We also experiment on the 1003 images of the standard dataset Judd (Judd et al., 2009), which can be considered non-iconic, although we have no statistics of the number of objects or their distribution in the images.

4.2 Candidate regions

In this section, we evaluate the coverage of our candidate regions. Initially, we obtain 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) (Carreira and Sminchisescu, 2012) 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 (Arbelaez et al., 2012), 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 (Arbelaez et al., 2012). But we achieve it with nearly half the number of combinations – 721 compared to 1,322 – and with a simpler and faster algorithm (4s per image in its Matlab prototype). In table (1) we refer to their segments as UCM-combs and to our candidate regions as CR-combs. (III) The Selective Search (Uijlings et al., 2013) 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 (Uijlings et al., 2013) present results for bounding boxes and not for regions.) See table (1) with the region-based IoU and recall results.

	all IoU	recall	# cands.	time
CPMC	59.6	57.6%	**150**	250s
UCM-combs	**77.0**	**80.0%**	1322	850s
Sel. search	67.8	66.1%	2100	**4s**
Our CR combs	**74.0**	**70.3%**	**721**	**4s**

Table 1: Region-based IoU (in %) comparison. CPMC (Carreira and Sminchisescu, 2012), UCM (Arbelaez et al., 2012), Sel. Search (Uijlings et al., 2013), and our CR – candidate regions. Boldface denotes the first and second best results.

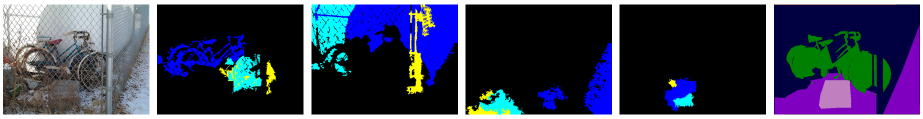


Figure 14: Left-to-right and top-to-bottom: Original image, top three segments for bike, wall, snow, rock, and ground truth. Note that the segments are good even for object classes that perform poorly overall (e.g., bike).

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 (Li et al., 2014) and Judd (Judd et al., 2009). We consider Pascal a more complicated case, based on the lower performance of the state-of-the-art methods. In Pascal our method outperforms the state of the art. On the Judd dataset only AWS (Garcia-Diaz et al., 2012) outperforms our method.

Our saliency model is part of a bottom-up visual processing framework. A possible drawback of using a multistep approach (segmentation, base-detail decomposition, saliency computation) is that if the initial segmentation has some imprecision, it can produce artifacts in the base-detail decomposition, leading to wrong saliency predictions.

In Fig. 15 we show a comparison of our Base-Detail Saliency (BDS) method, Adaptive Whitening Saliency (AWS, (Garcia-Diaz et al., 2012)), Image Signature (SIG, (Hou et al., 2012)) and L. Itti's original model (Itti, (Itti et al.,

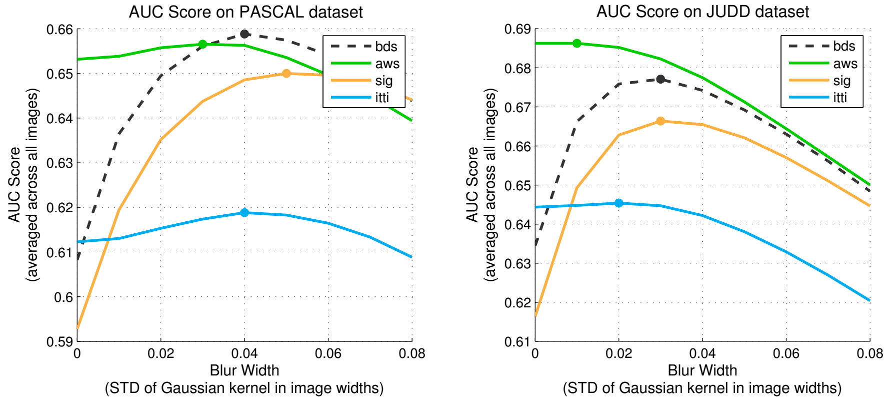


Figure 15: Bottom-up saliency performance. Left: Pascal-S dataset (Li et al., 2014). Right: Judd dataset (Judd et al., 2009). Approaches compared: Our Base-Detail Saliency (BDS), Adaptive Whitening Saliency (AWS, (Garcia-Diaz et al., 2012)), Image Signature (SIG, (Hou et al., 2012)), L. Itti's original model (Itti, (Itti et al., 1998)).

1998)). In Fig. 16 we show some examples for qualitative comparison between the results of the different algorithms.

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

Achanta, R., Shaji, A., Smith, K., Lucchi, A., Fua, P., and Susstrunk, S. (2012). SLIC superpixels compared to state-of-the-art superpixel methods. *TPAMI*,

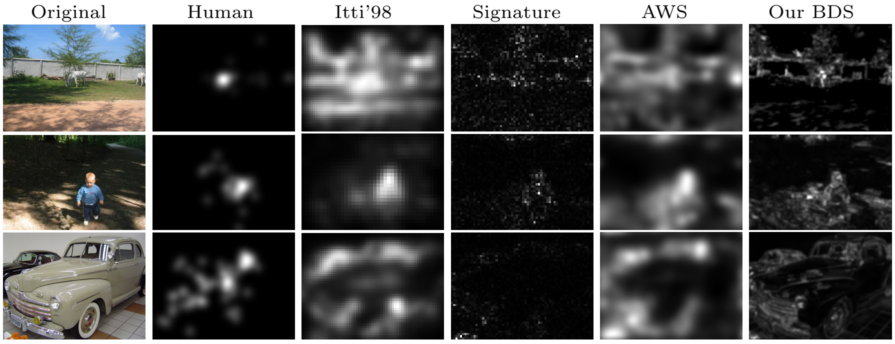


Figure 16: From left ro tight: Original; Human(fixations collected on 8 subjects with free-viewing task, first 3 seconds) (Li et al., 2014); Itti's original model (Itti et al., 1998); Spectral signature (Hou et al., 2012); AWS (Garcia-Diaz et al., 2012); Our Base-Detail Saliency (BDS) Bonev & Yuille.

34(11):2274–2282.

Alexe, B., Deselaers, T., and Ferrari, V. (2012). Measuring the objectness of image windows. *TPAMI*, 34(11):2189–2202.

Alpert, S., Galun, M., Brandt, A., and Basri, R. (2012). Image segmentation by probabilistic bottom-up aggregation and cue integration. *TPAMI*, 34(2):315–327.

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, pages 182–, Washington, DC, USA. IEEE Computer Society.

Arbelaez, P., Hariharan, B., Gu, C., Gupta, S., and Malik, J. (2012). Semantic segmentation using regions and parts. In *CVPR*.

Arbelaez, P., Maire, M., Fowlkes, C., and Malik, J. (2011). Contour detection and hierarchical image segmentation. *TPAMI*, 33(5):898–916.

Bae, S., Paris, S., and Durand, F. (2006). Two-scale tone management for photographic look. *ACM Trans. Graph.*, 25(3):637–645.

Barron, J. T. and Malik, J. (2012). Color constancy, intrinsic images, and shape estimation. *ECCV*.

Barrow, H. G. and Tenenbaum, J. M. (1978). Recovering intrinsic scene characteristics from images. Technical Report 157, AI Center, SRI International, 333 Ravenswood Ave., Menlo Park, CA 94025.

Bonev, B. and Yuille, A. L. (2014). A fast and simple algorithm for producing candidate regions. In *European Conference on Computer Vision (ECCV 2014)*.

Borji, A., Cheng, M., Jiang, H., and Li, J. (2014). Salient object detection: A survey. *CoRR*, abs/1411.5878.

Borji, A. and Itti, L. (2013). State-of-the-art in visual attention modeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(1):185–207.

Borji, A., Sihite, D. N., and Itti, L. (2013). Objects do not predict fixations better than early saliency: A re-analysis of einhäuser et al.'s data. *Journal of Vision*, 13(10).

Bradley, C., Abrams, J., and Geisler, W. S. (2014). Retina-v1 model of detectability across the visual field. *Journal of vision*, 14(12):22.

Carreira, J. and Sminchisescu, C. (2012). Cpmc: Automatic object segmentation using constrained parametric min-cuts. *TPAMI*, 34(7):1312–1328.

Einhäuser, W., Spain, M., and Perona, P. (2008). Objects predict fixations better than early saliency. *Journal of Vision*, 8(14).

Everingham, M., Van Gool, L., Williams, C. K., Winn, J., and Zisserman, A. (2010). The PASCAL visual object classes (VOC) challenge. *International journal of computer vision*, 88(2):303–338.

Farbman, Z., Fattal, R., Lischinski, D., and Szeliski, R. (2008). Edge-preserving decompositions for multi-scale tone and detail manipulation. *ACM Trans. Graph.*, 27(3):67:1–67:10.

Felzenszwalb, P. F. and Huttenlocher, D. P. (2004). Efficient graph-based image segmentation. *IJCV*, 59(2):167–181.

Galun, M., Sharon, E., Basri, R., and Brandt, A. (2003). Texture segmentation by multiscale aggregation of filter responses and shape elements. ICCV '03, pages 716–.

Garcia-Diaz, A., Leborán, V., Fdez-Vidal, X. R., and Pardo, X. M. (2012). On the relationship between optical variability, visual saliency, and eye fixations: A computational approach. *Journal of Vision*, 12(6):1–22.

Geman, S. and Geman, D. (1984). Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *IEEE Trans. Pattern Anal. Mach. Intell.*, 6(6):721–741.

Gollisch, T. and Meister, M. (2010). Eye smarter than scientists believed: neural computations in circuits of the retina. 65(2):150–164.

Golub, G. H. and Van Loan, C. F. (2012). *Matrix computations*, volume 3. JHU Press.

Gonzalez, R. C., Woods, R. E., and Eddins, S. L. (2004). Digital image processing using matlab. *Upper Saddle River, N. J: Pearson Prentice Hall.*

Gorelick, L. and Basri, R. (2009). Shape based detection and top-down delineation using image segments. *Int. J. Comput. Vision*, 83(3):211–232.

Horn, B. K. P. and Brooks, M. J. (1986). The variational approach to shape from shading. *Computer Vision, Graphics, and Image Processing*, 33(2):174–208.

Hou, X., Harel, J., and Koch, C. (2012). Image signature: Highlighting sparse salient regions. *TPAMI, IEEE*, 34(1):194–201.

Humayun, A., Li, F., and Rehg, J. M. (2014). RIGOR: Reusing Inference in Graph Cuts for generating Object Regions. In *Computer Vision and Pattern Recognition (CVPR), Proceedings of IEEE Conference on*. IEEE.

Itti, L., Koch, C., and Niebur, E. (1998). A model of saliency-based visual attention for rapid scene analysis. *TPAMI, IEEE*, 20(11):1254–1259.

Jian Li, Martin Levine, X. A. and He, H. (2011). Saliency detection based on frequency and spatial domain analyses. In *Proc. BMVC*, pages 86.1–86.11. http://dx.doi.org/10.5244/C.25.86.

Judd, T., Ehinger, K., Durand, F., and Torralba, A. (2009). Learning to predict where humans look. In *ICCV*, pages 2106–2113. IEEE.

Land, E. H. (1977). The retinex theory of color vision.

Leclerc, Y. (1989). Image and boundary segmentation via minimal-length encoding on the connection machine. In *DARPA89*, pages 1056–1069.

LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. (1998). Gradient-based learning applied to document recognition. In *Proceedings of the IEEE*, volume 86, pages 2278–2324.

Leonenko, N., Pronzato, L., and Savani, V. (2008). A class of Rényi information estimators for multidimensional densities. *Ann. Statist.*, 36(5):2153–5182.

Li, J., Levine, M. D., An, X., Xu, X., and He, H. (2013). Visual saliency based on scale-space analysis in the frequency domain. *IEEE Trans. Pattern Anal. Mach. Intell.*, 35(4):996–1010.

Li, Y., Hou, X., Koch, C., Rehg, J. M., and Yuille, A. L. (2014). The secrets of salient object segmentation. In *CVPR*.

Marr, D. (1982). *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information.* Henry Holt and Co., Inc., New York, NY, USA.

Mottaghi, R., Chen, X., Liu, X., Fidler, S., Urtasun, R., and Yuille, A. (2014). The role of context for object detection and semantic segmentation in the wild. In *CVPR*.

Russ, J. C. and Woods, R. P. (1995). The image processing handbook. *Journal of Computer Assisted Tomography*, 19(6):979–981.

Shapley, R. and Enroth-Cugell, C. (1984). Visual adaptation and retinal gain controls. *Progress in retinal research*, 3:263–346.

Todorovic, S. and Ahuja, N. (2008). Region-based hierarchical image matching. *IJCV*, 78(1):47–66.

Tomasi, C. and Manduchi, R. (1998). Bilateral filtering for gray and color images. In *Computer Vision, 1998. Sixth International Conference on*, pages 839–846. IEEE.

Tu, Z., Zhu, S.-C., and Shum, H.-Y. (2001). Image segmentation by data driven markov chain monte carlo. In *Computer Vision, 2001. ICCV 2001. Proceedings. Eighth IEEE International Conference on*, volume 2, pages 131–138 vol.2.

Uijlings, J. R. R., van de Sande, K. E. A., Gevers, T., and Smeulders, A. W. M. (2013). Selective search for object recognition. *International Journal of Computer Vision*, 104(2):154–171.

Woodham, R. J. (1980). Photometric method for determining surface orientation from multiple images. *Optical engineering*, 19(1):191139–191139.

Xu, C., Xiong, C., and Corso, J. J. (2012). Streaming hierarchical video segmentation. In *ECCV*.

Yuan, L. and Sun, J. (2012). Automatic exposure correction of consumer photographs. In Fitzgibbon, A. W., Lazebnik, S., Perona, P., Sato, Y., and Schmid, C., editors, *ECCV (4)*, volume 7575 of *Lecture Notes in Computer Science*, pages 771–785. Springer.

Zhaoping, L. (2003). V1 mechanisms and some figure-ground and border effects. *Journal of Physiology*, 97(1):503–515.

Zhaoping, L. (2014). *Understanding Vision: Theory, Models, and Data*. Oxford.

Zhu, L., Chen, Y., Lin, Y., Lin, C., and Yuille, A. (2012). Recursive segmentation and recognition templates for image parsing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(2):359–371.

Zhu, S. C. and 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.

Zhu, Y., Zhang, Y., and Yuille, A. (2014). Single image super-resolution using deformable patches. In *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, pages 2917–2924.

Zitnick, C. L. and Dollár, P. (2014). Edge boxes: Locating object proposals from edges. In *ECCV*.