Vision Research 120 (2016) 93–107

Contents lists available at ScienceDirect

Vision Research

journal homepage: www.elsevier.com/locate/visres

A systematic comparison between visual cues for boundary detection

David A. Mély [a,b], Junkyung Kim [a,b], Mason McGill [a,b], Yuliang Guo [a,c], Thomas Serre [a,b,d,*]

[a] Brown University, Providence, RI 02912, United States
[b] Department of Cognitive, Linguistic and Psychological Sciences, United States
[c] Department of Engineering, United States
[d] Brown Institute for Brain Science, United States

ARTICLE INFO

Article history:
Received 21 August 2014
Received in revised form 17 November 2015
Accepted 17 November 2015
Available online 2 March 2016

Keywords:
Boundary
Contour
Segmentation
Grouping
Early vision
Primary visual cortex
Natural scenes

ABSTRACT

The detection of object boundaries is a critical first step for many visual processing tasks. Multiple cues (we consider luminance, color, motion and binocular disparity) available in the early visual system may signal object boundaries but little is known about their relative diagnosticity and how to optimally combine them for boundary detection. This study thus aims at understanding how early visual processes inform boundary detection in natural scenes. We collected color binocular video sequences of natural scenes to construct a video database. Each scene was annotated with two full sets of ground-truth contours (one set limited to object boundaries and another set which included all edges). We implemented an integrated computational model of early vision that spans all considered cues, and then assessed their diagnosticity by training machine learning classifiers on individual channels. Color and luminance were found to be most diagnostic while stereo and motion were least. Combining all cues yielded a significant improvement in accuracy beyond that of any cue in isolation. Furthermore, the accuracy of individual cues was found to be a poor predictor of their unique contribution for the combination. This result suggested a complex interaction between cues, which we further quantified using regularization techniques. Our systematic assessment of the accuracy of early vision models for boundary detection together with the resulting annotated video dataset should provide a useful benchmark towards the development of higher-level models of visual processing.

1. Introduction

Natural scenes constitute a rich source of visual information, carried by a variety of visual cues. Previous work has shown that our visual system does indeed rely on a combination of cues to solve different visual tasks including orientation (Cavanagh, 1992) and depth (Young, Landy, & Maloney, 1993; Johnston, Cumming, & Landy, 1994; Tassinari, Domini, & Caudek, 2008) analysis, biological motion recognition (Thurman & Lu, 2013) and object detection (Vuong, Hof, Bülthoff, & Thornton, 2006).

Because image properties are expected to differ markedly across object boundaries, boundary detection is a prime testbed to study the diagnosticity of individual cues and of their combinations. For instance, borders defined by both chromatic and luminance cues are more visible than those defined by only one of these cues, with each cue making independent contributions (Frome, Buck, & Boynton, 1981). Beyond the pairing of chromatic and luminance cues, there exist complex interactions between all pairings of luminance, motion, color and texture cues (Rivest & Cavanagh, 1996). In the context of texture-defined boundaries, an ideal observer model of cue combination accounts well for annotators' judgment of the relative location of two edges (Landy & Kojima, 2001).

Consistent with the results from these psychophysical studies, neurophysiology studies have shown the existence of cue-independent boundary-selective neurons in higher order visual areas in both the ventral (Sary, Vogels, & Orban, 1993) and dorsal (Albright, 1992; Geesaman & Andersen, 1996) streams of visual processing in cortex. Some studies have reported the existence of such cue-independent neurons in early stages of visual processing in areas V1 and V2 (Leventhal, Thompson, Liu, Zhou, & Ault, 1995; Leventhal, Wang, Schmolesky, & Zhou, 1998; Sincich & Horton, 2005).

Overall, using controlled artificial stimuli, previous work has demonstrated our visual system's ability to efficiently combine visual cues for the detection of boundaries. However, to date,

* Corresponding author at: Brown University, Providence, RI 02912, United States.
E-mail addresses: david_mely@brown.edu (D.A. Mély), junkyung_kim@brown.edu (J. Kim), mmcgill@caltech.edu (M. McGill), yuliang_guo@brown.edu (Y. Guo), thomas_serre@brown.edu (T. Serre).

http://dx.doi.org/10.1016/j.visres.2015.11.007
0042-6989/

relatively little is known about the diagnosticity of these various cues in natural environments and how they should be optimally combined. One promising research direction is to try to characterize the relevant environmental constraints for the detection of boundaries. In particular, the past decade of research has witnessed a rapid growth in the use of statistical methods towards natural scene analysis to better understand what types of visual cues provide useful information (see Simoncelli & Olshausen, 2001; Geisler, 2008 for reviews).

Early work has shown that there exists pairwise statistical dependencies between nearby pixels, and that these dependencies are consistent with the Gestalt principles of co-linearity and parallelism (Krüger, 1998). Further work examined second-order spatial statistics and found long-range correlations that adhered to the geometric principle of co-circularity (Sigman, Cecchi, Gilbert, & Magnasco, 2001). However, the statistical analysis in both of these studies was not limited to object boundaries but included lower-level edges, possibly part of textures and shading flows.

Two important studies expanded on this line of work by limiting their statistical analyses to hand-annotated boundaries in natural images. Geisler et al. measured the pairwise statistics of edge elements as a function of their geometry and contrast polarity (Geisler, Albrecht, Crane, & Stern, 2001; Geisler & Perry, 2004). When considering the geometrical and contrast relationship between one reference and another element, for any given distance between the two elements, it was found that the most likely geometrical relationship was one consistent with approximate co-circularity (Field, Hayes, & Hess, 1993; Kapadia, Ito, Gilbert, & Westheimer, 1995; Adini, Moses, & Ullman, 1997; Li, 1998; Geisler et al., 2001). Exploiting this co-circularity principle was shown to yield better detection accuracy (Li, 1998; Ross, Badcock, & Hayes, 2000; VanRullen, Delorme, & Thorpe, 2001). Elder and Goldberg further extended the approach by modeling the sequence of elements along a boundary using a Markov chain, and considering arbitrary pairs of tangents on the boundary to compute pairwise statistics of edge elements (Elder & Goldberg, 2002).

Beyond grayscale images, the study of the spatio-chromatic structure of natural scenes suggests that luminance and chromatic edges are not independent of each other (Fine, MacLeod, & Boynton, 2003). In fact, they tend to co-occur around boundaries salient to human perception (Zhou & Mel, 2008) and are linked to higher-order statistical dependencies (but see also Hansen & Gegenfurtner, 2009). Several researchers have suggested that incorporating mechanisms of divisive normalization (which is a form of nonlinear gain control found in cortex (Carandini & Heeger, 2012)) over the outputs of edge operators would effectively reduce these correlations (Simoncelli & Schwartz, 1999; Zhou & Mel, 2008; Ramachandra & Mel, 2013; see also Zetzsche, 2001 for a review).

More recently, a study assessed the diagnosticity of texture, luminance and color cues for the detection of boundaries in images of close-up foliage, an ecologically important component of primates' natural environment (Ing, Wilson, & Geisler, 2010). Various statistical classifiers were trained on multiple image measurements (including luminance and chromaticity computed over individual image patches as well as the difference along these dimensions between pairs of patches). Participants judged whether pairs of image patches sampled from the same scenes at some spatial separation belonged to the same physical surface. An ideal classifier, which combined optimally all image measurements, was found to agree well with human psychophysics data for the same task.

The role of various visual cues for image segmentation has also been studied from the perspective of computer vision. One can distinguish between two general approaches to image segmentation. Edge-based approaches typically rely on the direct detection of boundary elements by coinciding edge detectors which are then grouped together to form boundaries. Early approaches including the popular Canny, Sobel and Prewitt detectors, have initially focused on local gradient computations (see Bowyer, Kranenburg, & Dougherty, 2001 for a comparison between representative approaches). These edge-based approaches were later extended to color images using color-edge operators (Koschan & Abidi, 2005; Tamrakar & Kimia, 2007).

Region-based approaches, on the other hand, aim at partitioning an image by assessing the low-level coherence of individual visual cues (such as luminance, color, texture, or motion attributes) to sequentially partition an image into regions (*e.g.*, Felzenszwalb & Huttenlocher, 2004; Cremers, Rousson, & Deriche, 2006; Moore, Prince, Warrell, Mohammed, & Jones, 2008; Levinshtein et al., 2009; Veksler, Boykov, & Mehrani, 2010; Achanta et al., 2012). A prominent example is the *Pb* system (Martin, Fowlkes, & Malik, 2004) and its derivatives *mPb* and *gPb* (Arbelaez, Maire, Fowlkes, & Malik, 2010), which yielded state-of-the-art results on natural scenes. It extends a popular class of biologically-motivated algorithms for boundary detection based on the *filter–rectify–filter* model (*e.g.*, Wermser & Liedtke, 1982; Voorhees & Poggio, 1988; Bovik, Clark, & Geisler, 1990; Malik & Perona, 1990; Caelli, 1993), which are built on the rectified outputs of filter banks such as Gabor wavelets, Gaussian derivatives or steerable filters that coarsely mimic processing by orientation-selective cells found in the primary visual cortex (V1) (Hubel & Wiesel, 1962).

In this approach, a unique pattern of activation across all filters is considered to be a distinct texture, and boundaries are detected when two neighboring regions of the image are composed of different textures. The key operation is an additional filtering stage, based on the χ^2 operator (also referred to as the oriented gradient operator), which divides a local circular neighborhood in the image into two halves along some orientation and measures the difference between empirical distributions of the cue values in either of the two halves. This difference is computed using a χ^2 histogram distance between binned estimates of the distribution of values taken by the given cue.

More recently, the *Pb* approach was extended to combine static boundary cues with low-level motion cues computed from a motion-gradient channel as well as motion cues derived from optical flow (Sundberg, Brox, Maire, Arbaelez, & Malik, 2011) (see also *e.g.*, Wang & Adelson, 1994, 1995, 1998, 1998, 2000 for earlier representative work offering evidence that motion cues contain figure-ground and depth ordering information in addition to boundary information).

Depth cues are also useful for boundary detection and image segmentation. For instance, disparity information estimated from stereo cameras can be combined with edges computed from luminance and/or chrominance values (Woo, Kim, & Iwadate, 2000; Gelautz & Markovic, 2004). More recently, because of the increasing availability of RGB-D data from depth sensors, several studies have focused on range data to detect boundaries (*e.g.*, Silberman, Hoiem, Kohli, & Fergus, 2012; Ren, Bo, & Fox, 2012; Gupta, Arbel, & Malik, 2013). In particular, the *Pb* framework was extended to include depth, convex normal, and concave normal gradient channels (Gupta et al., 2013).

The diagnosticity of particular cues for object boundary detection (whether located on or off a boundary) may depend on their extended visual context. Learning-based approaches have proven successful in constructing mid-level visual representations able to incorporate said context (Dollar & Belongie, 2006; Lim, Zitnick, & Dollar, 2013). Such approaches are a hybrid between edge-based and region-based boundary detection.

Overall, previous work in early vision has demonstrated that the combination of low-level visual cues, including color, motion,

depth and luminance, enables more accurate segmentation of natural scenes. However, the diagnosticity of early visual cues alone or in combination has not been studied systematically. This study thus aims at gaining a deeper understanding of how early visual processes inform boundary detection in natural scenes.

To address this question, we collected a video dataset, which consists of short color stereo video sequences collected with a consumer-grade camera. The dataset was manually annotated to provide a ground-truth for the locations of physical boundaries. Inspired by the notion of elemental measurements in early vision put forth by Adelson and Bergen (1991), we further implemented an integrated approach to derive such visual elements across multiple cues (luminance, color, motion and stereo) in a systematic way[1]. Here, we considered plausible scenarios to extract boundary signals from these visual elements: an edge-based (first-order) approach where the responses of the filters to different cues are used directly to detect object boundaries, and a region-based (second-order) approach based on the aforementioned *Pb* framework applied to the proposed visual elements computed over different cues. We trained multiple machine learning classifiers on the various cues—both in isolation and in combination—and assessed the accuracy of the resulting classifiers for the detection of boundaries.

2. Materials and methods

2.1. Video collection

We built a rich video dataset composed of short binocular video sequences of natural scenes using a consumer-grade (Fujifilm) stereo camera. We considered a variety of locations (from university campuses to street scenes and parks) and seasons to minimize possible biases. We attempted to capture more challenging scenes for boundary detection by framing a few dominant objects in each shot under a variety of viewing angles, distances, and lighting conditions. All sequences were recorded by a moving observer (either on foot or in a motorized vehicle). Additionally, about 60% of the frames included objects moving in the scene (including zoo animals, pets, pedestrians and cars). The dataset contains 100 scenes, each consisting of a short (10-frame) stereo (left and right views) color sequence. Each sequence was sampled at a rate of 30 frames per second. Each frame had a resolution of 1280 by 720 pixels. Representative frames are shown in Fig. 1.

2.2. Ground-truth annotations

Direct measurements of ground-truth information for the location of physical boundaries would be extremely challenging in natural scenes; hence a common shortcut is to rely on manual segmentation by human observers (Geisler et al., 2001; Elder & Goldberg, 2002; Martin et al., 2004). A common assumption is that, given enough time and the ability to zoom in and out of an image, annotators can accurately recover boundaries from the environment, *e.g.*, ones due to differences between the material properties of two physically distinct surfaces. As we will discuss later, the relatively high inter-subject consistency resulting from this procedure suggests that manual boundary annotation can provide useful ground-truth information.

Here, we computed ground-truth boundary information by collecting two sets of manual annotations for the last frame of the left stereo view for each individual video sequence. Thereafter, we refer to the main set of annotations as the *boundary annotations*. Unless specified otherwise, this is the set used for most analyses in the paper. Hand-segmentation was performed by paid undergraduate students ($n = 5$) at Brown University (Providence, RI). We wrote custom software to enable manual annotations within a web browser. Annotators were not limited in the amount of time they had available to complete the task. They were paid by time spent annotating boundaries, to the condition that they annotate the entirety of the dataset.

Segmentation involved annotating the contours defining the boundaries of each object's visible surface regions. We gave all annotators the same basic instructions as were provided for the Berkeley dataset (see Martin et al., 2004 for details): "*You will be presented a photographic image. Divide the image into some number of segments, where the segments represent "things" or "parts of things" in the scene. The number of segments is up to you, as it depends on the image. Something between 2 and 30 is likely to be appropriate. It is important that all of the segments have approximately equal importance*".

Representative annotations are shown in Fig. 2. Additional samples can be visualized by browsing the database at http://serre-lab.clps.brown.edu/resource/multicue. Fig. S1 shows a histogram illustrating the distribution of the number of labeled pixels in the images.

To compute an estimate of the inter-subject agreement, we conducted a leave-one-annotator-out procedure, where each annotator was considered exactly once as a cue, and evaluated against the union of the rest of the annotators. In a sense, the latter set of all remaining annotators is tested like any other ground-truth, using the same metric (the F-measure; see below). We found manual annotations to be highly consistent despite the challenges associated with the collection of such annotations arising, *e.g.*, because of inherent differences in the strategies used by the observers (see Guo & Kimia, 2012 for a discussion). We report the following F-measure for inter-subject agreement: $F = .76 \pm 0.017$, which is comparable to the $F = .80$ reported in Martin et al. (2004) for the Berkeley dataset.

Subsequently, we collected a set of lower-level *edge annotations* (inter-subject agreement: $F = .75 \pm 0.024$), by asking participants to annotate all edges and boundaries in the image; in this case, tracing contours that did not correspond to "things" or physical boundaries such as shadows was encouraged. This new set of lower-level annotations is reminiscent of Guo and Kimia (2012) but extends their approach to binocular video sequences. Fig. 2 shows the difference between the two sets of annotations on representative scenes; see also Fig. S1.

2.3. Accuracy measure

To evaluate the accuracy of the different cues against ground-truth boundary annotations, we used the Berkeley evaluation software (Arbelaez et al., 2010). This evaluation pipeline, which is standard in the field, matches machine-produced boundaries against hand annotations while allowing a reasonable amount of slack, as opposed to a naive pixel-wise difference between the two. This is essential for proper evaluation of the different cues as a naive scheme could potentially over-penalize the detection of boundaries at slightly offset locations. It is not reasonable to expect an algorithm to retrieve perfect boundary localization since the human annotations display such variability as well.

The evaluator returns a precision-recall curve for any boundary classifier output. Here we report the F-measure, a standard accuracy measure that captures an even balance between precision and recall. Other standard measures include precision and recall from signal detection theory, as well as the area under the precision-recall curve and will be given in the Supplementary Information.

[1] Similar visual elements based on oriented filters also form the backbone of the Berkeley *Pb* segmentation system, which exhibits state-of-the-art accuracy (Martin et al., 2004; Arbelaez et al., 2010).

Fig. 1. Representative frames sampled from the dataset.

2.4. Machine learning classifiers

The *Scikit-Learn* library (Pedregosa et al., 2011) was used to evaluate the accuracy of the various cues and their combination. We used a L^2-norm regularized logistic classifier, which was shown to perform on par with more complex classifiers for this problem (Martin et al., 2004). The regularization parameter was selected using stratified 3-fold cross-validation. All accuracy metrics were averaged over 3 random splits of the training/test data with 80 images for training and 20 for test. All cues were optimized, either individually or in combination, using the same pipeline, unless otherwise specified. Each classifier was trained on one million samples.

The approach in Arbelaez et al. (2010) uses coordinate ascent on the F-measure as an objective function. It is computationally prohibitive as it involves computing the F-measure over the whole training set for every cycle of the learning phase. Thus, we departed from the original paradigm described in Arbelaez et al. (2010) and used a pixel-wise error measure for training but kept the F-measure for final evaluation.

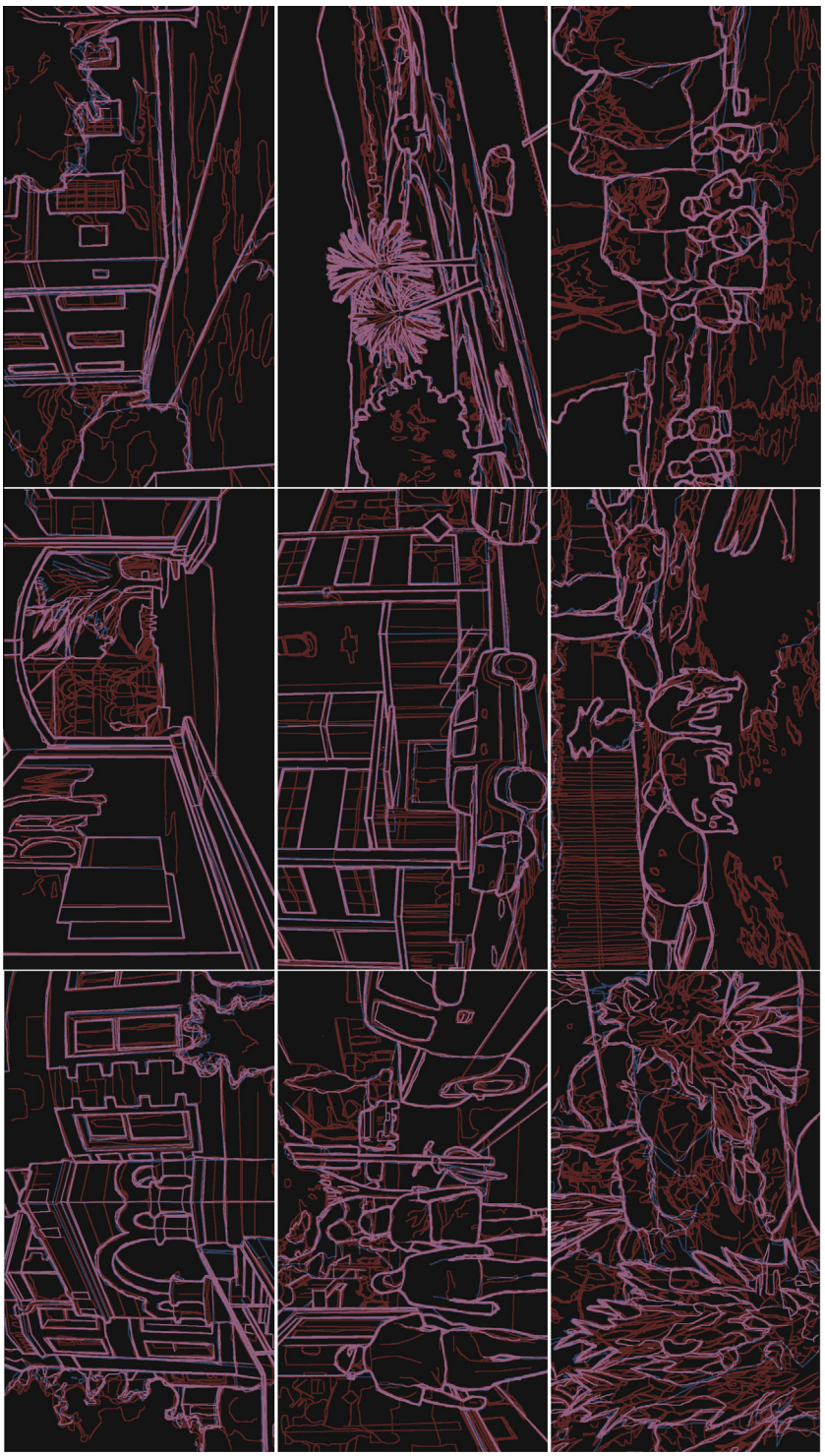

Fig. 2. Boundary annotations for the sample scenes shown in Fig. 1. Blue lines correspond to our default set of annotations, *i.e.*, higher-level, object-centric boundary annotations. Red lines correspond to (finer-grained) lower-level edge annotations collected as a control (see Methods for details on how these annotations were collected). Pink lines correspond to annotations that are common to both sets. Annotations have been dilated for display purposes.

The visual features provided to each classifier were typically a few pixels wider than the boundary annotations, which was problematic since it was likely to respond strongly on the pixels immediately next to the annotation that were by definition labeled as non-boundary. Hence, we did not incorporate pixels that were within five pixels to the closest ground-truth boundary in the training set.

3. Computational model

All software used for the computation of visual features was written in *Python*, and used Nvidia's cuDNN bindings to accelerate the computations on graphical processing units (GPU). We used Seaborn/Matplotlib for data visualization. The data set and

source code is made available at http://serre-lab.clps.brown.edu/resource/multicue.

All experiments ran on Brown's large computing cluster at the Center for Computation and Visualization (CCV) with >500 Teraflops of computing power. Feature computation across all visual cues took about 100 computing nodes/processes, 64 GB of memory each, for a couple of days. Training individual classifiers on individual nodes required 256 GB of RAM for at most a week. Assessing the accuracy of cues using the benchmarking software from (Arbelaez et al., 2010) required about 200 processes for a week.

3.1. Luminance

Our starting point was a standard battery of Gabor filters, which, along with derivative-of-Gaussian filters, constitutes the backbone of many boundary detection systems (Malik & Perona, 1990; Martin et al., 2004; Arbelaez et al., 2010). In our implementation, we considered 3 filter sizes (or scales): 7-by-7, 13-by-13, and 19-by-19 pixels. For each filter size s, the standard deviation of the Gaussian envelope was fixed at $s/3$. Additionally, we considered, for each filter size s, 3 spatial frequencies for the harmonic component of the Gabor filter: $s/4$, $s/2$, and s. We considered 8 equally-spaced orientations for each combination of filter size and spatial frequency. We added one center-surround channel for each filter size, for a total of $N = 3 * (3 * 8 + 1) = 75$ filters for the luminance model. Then the orientation energy was computed directly from the outputs of quadrature pairs of filters with a phase difference of $\pi/2$. Lastly, we considered a normalization step over multiple orientations (see below). The use of Gabor filters, the energy model and normalization are standard in studies of natural boundary statistics (*e.g.*, Krüger, 1998; Geisler et al., 2001; Geisler, Perry, & Ing, 2008; Sigman et al., 2001).

Earlier studies relied on fixed-scale filters (Krüger, 1998; Sigman et al., 2001; Geisler et al., 2001) (but see also Elder & Goldberg, 2002). Using filters that vary over a range of scales and spatial frequencies was necessary to capture variations of the image power spectrum across natural scenes. It was also found experimentally that the addition of multiple scales and spatial frequencies yielded a moderate improvement in boundary detection accuracy (contrary to what was found in Arbelaez et al. (2010) which only considered multiple scales; results not shown). We contend the cause for this improvement is the inclusion of multiple spatial frequencies in addition to filter sizes, which allows some filters to be selected for their robust responses on richly-textured surfaces, and others to be leveraged for their preferential responses to boundaries (see below). We next describe how we extended this basic luminance channel to color, stereo and motion. An overview of the proposed multi-cue approach is shown in Fig. 3.

3.2. Color

Several functions have been attributed to color vision, from helping to find edible food (Dominy & Lucas, 2001; Regan et al., 1998) to discriminating emotional states, socio-sexual signals and threat displays on the skin of conspecifics (Changizi, Zhang, & Shimojo, 2006) as well as facilitating scene and object recognition (Gegenfurtner & Rieger, 2000; Wurm, Legge, Isenberg, & Luebker, 1993) and guiding visual searches in real-world scenes (Ehinger & Brockmole, 2008) in addition to boundary detection (Lotto, Clarke, Corney, & Purves, 2011). Psychophysical studies have shown that chromatic mechanisms are spatially tuned for both orientation and spatial frequency (Humanski & Wilson, 1993; Mullen & Losada, 1999), supporting the idea that these channels have an early cortical representation (Shapley & Hawken, 2011).

Here, we considered the early vision color model developed by our group (Zhang, Barhomi, & Serre, 2012). This framework extends the base Gabor (plus center-surround) filter pyramids to processing along three chromatic opponent axes: *R–G* (red versus green), *R–C* (red versus cyan), and *B–Y* (blue versus yellow), in addition to a luminance-based *Wh–Bl* channel. We included both single-opponent (SO) and double-opponent (DO) color channels as described in Zhang et al. (2012). The model is similar to the approach described in Johnson, Kingdom, and Baker (2005) with the SO and DO stages closely related to first- and second-order channels. Differences include an additional rectification stage at the SO level (consistent with physiology) and a divisive normalization stage proposed in Zhang et al. (2012) (see later), which are key to robust boundary detection accuracy. In particular, the additional normalization stage was found to be important both to reduce spurious edge responses (Zhang et al., 2012) and to account for psychophysical data of color similarity ratings (Zhang, Mély & Serre, manuscript in preparation).

As shown in Zhang et al. (2012), SO units are strongly modulated by chromatically opponent interactions between their center and their surround (*e.g.*, the response of the unit is facilitated by blue light in the center and suppressed by yellow light in its surround). However, these units are only weakly tuned to orientation. As speculated in Hurlbert (1989) and demonstrated in Zhang et al. (2012), such units respond mainly to surfaces. Conversely, DO units are selective for a preferred chromatic opponent pair and sharply tuned for orientation and frequency and respond mainly to edges defined by (equiluminant) color contrast. While the existence of DO cells was initially subject to debate, it is now relatively well established (see Shapley & Hawken, 2011 for a review).

We considered complex DO units that are invariant to a chromatic contrast reversal (*e.g.*, red-green versus green–red). Their responses are obtained by computing the energy of the two filters within a preferred color-opponent pair. The corresponding units are able to capture a wide range of equiluminant chromatic edges and textures that would otherwise remain undifferentiated by a purely luminance-based model.

3.3. Motion

Motion is a powerful cue to figure-ground segmentation. Many biological organisms rapidly segregate camouflaged objects that are set in motion (Segaert, Nygård, & Wagemans, 2009; Uttal, Spillmann, Stürzel, & Sekuler, 2000) from their background. Infants as early as 3 month old are already able to detect kinetic boundaries defined exclusively by motion cues (Kaufmann-Hayoz, Kaufmann, & Stucki, 1986).

Here, we extended the basic oriented (luminance) filters described above from the spatial domain to the spatio-temporal domain by constructing spatio-temporal (3D) filters obtained by drifting a base 2D filter orthogonally to its preferred orientation along a (third) time dimension. This is equivalent to adding a time-dependent periodic phase shift in the harmonic component of the Gabor wavelet and multiplying the Gaussian term with a temporal envelope term. The inverse of this period parameter is also known as the temporal frequency, to which early motion-sensitive cells are known to be tuned (see Bradley & Goyal, 2008 for review). For a given range of temporal frequencies, the direction of drift is entirely determined by the filter's orientation preference such that a typically large oriented element will always be perceived to be drifting orthogonally to its direction by a filter with the corresponding orientation preference (this is known as the aperture problem; see Bradley & Goyal, 2008). This is also consistent with what is known about the early stages of motion perception (Bradley & Goyal, 2008). In our implementation, the sign of temporal frequency corresponded to coherent motion in either rightward or leftward direction.

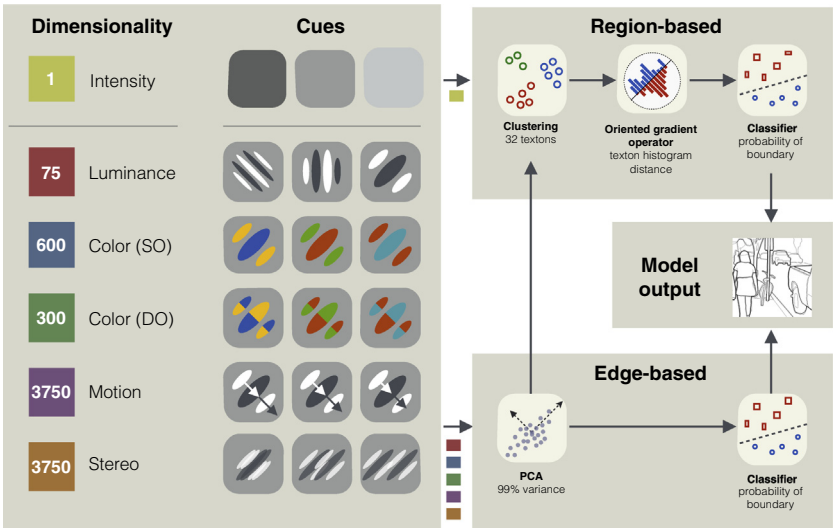

Fig. 3. Overview of the approach and the early visual cues considered. Principal component analysis (99% variance explained) is first used as a pre-processing for all cues except intensity (which unlike the others is based on raw pixel value instead of filter response) for a fair comparison between visual representations that differ greatly in their dimensionality. In the edge-based approach, filter responses are passed directly to a classifier trained for boundary detection. In the region-based approach, a mid-level visual representation is first learned to represent texture before it is passed to a classifier. The color code shown on the left for each individual cue is consistent throughout the paper. Dimensionality for different cues was computed as follows (sc. = scales/filter sizes, sf. = spatial frequencies, or. = orientations, cs. = center-surround.): lum. **75** = 3 sc. × (3 sf. × 8 or. + 1 cs.); so. **75** × 8 opponent channels; do. **75** × 4 opponent channels; mot. **75** × 50 temporal frequencies; ste. **75** × 50 binocular disparities.

The center-surround luminance filter channel was extended to the time domain by considering temporal modulation to detect local variations in image brightness devoid of coherent motion (also known as "flicker"). Together, following (Derpanis & Wildes, 2009), the spatio-temporal channel formed a basis covering the spatio-temporal frequency domain capable of representing dynamic textures. Alternative models of early motion processing have also been proposed, which include three-dimensional Gaussian derivative kernels (Simoncelli & Heeger, 1998).

3.4. Stereo

The model used for the computation of early stereo cues is part of an extended model developed by our group (Kim, Mély & Serre, manuscript in preparation) based on the disparity energy model (Ohzawa, DeAngelis, & Freeman, 1990; Ohzawa, 1998). In the model, the left and right views from each stereo scene are processed with binocular filters, *i.e.*, pairs of identical Gabor filters (*viz.*, each pair shares the same orientation, spatial frequency and phase) — one for each view. One of the filters is horizontally shifted with respect to the other in image coordinates; the displacement between the two filters is binocular disparity. For each pair, the monocular filter responses undergo summation followed by a rectified squaring operation. Surfaces located in front of the fixation plane have positive disparities, whereas those located behind it have negative disparities, with the magnitude of disparity increasing away in depth from the fixation plane. Thus, stereo cues may reveal boundaries between two image regions located at different depths but otherwise alike in visual properties.

3.5. Divisive normalization

In general, the raw outputs of the individual cue channels described above tend to be quite noisy. As mentioned below the disparity model is plagued by false matches between right and left features and it often fails in complex natural scenes (Kim, Mély & Serre, in preparation). Moreover, early vision models based on linear filters are typically characterized by statistical dependencies between filter outputs within any single cue, even when such filters are designed to maximize independence (Wainwright, Schwartz, & Simoncelli, 2002).

This is particularly problematic given that each cue independently considers several properties (*e.g.*, double-opponent color encompasses all possible combinations for scale/spatial frequency, chromatic contrast, and orientation), potentially resulting in undesirable statistical dependencies. Here, we consider divisive normalization, a computation known to alleviate this problem (Simoncelli & Schwartz, 1999); it is related to the eponymous form of inhibition in the visual cortex (see Carandini & Heeger, 2012 for a review).

Thus, we implemented divisive normalization over the appropriate pool of units: for each cue, we indexed each unit $X^{i,j}_{\theta_1,\theta_2,\ldots,\theta_n}$ by its respective preferred values along each of the sensory axes relevant to that cue, $\theta_1, \theta_2, \ldots, \theta_n$, as well as its position with respect to image coordinates (i, j). The unit gets normalized by all unit activities with the same preferred tuning except for one:

$$\overline{X}^{i,j}_{\theta_1,\theta_2,\ldots,\theta_n} = \frac{{X^{i,j}_{\theta_1,\theta_2,\ldots,\theta_n}}^{p}}{\left(\sigma^2 + \sum_{\phi_1} {X^{i,j}_{\phi_1,\theta_2,\ldots,\theta_n}}^{q} + \ldots + \sum_{\phi_n} {X^{i,j}_{\theta_1,\theta_2,\ldots,\phi_n}}^{q}\right)^{r}},$$

where σ^2 is a constant to avoid division by zero, p, q and r are arbitrary exponents and $\overline{X}$ is the normalized output. Such a scheme results in lower dependencies across each sensory axis for each cue. In object recognition, a simplified version of this model leads to very significant gains in accuracy with the proper normalization (Jarrett, Kavukcuoglu, Ranzato, & LeCun, 2009; Zhang et al., 2012).

As for binocular disparity, the literature suggests the existence of suppressive mechanisms between pairs of binocular inputs of opposite phases (corresponding to a phase difference of 180°) (Tanabe, Haefner, & Cumming, 2011). In practice, we have found that implementing such interactions via the divisive normalization circuit described above, whereby one binocular input at one (position) disparity is inhibited by the outputs corresponding to the same (position) disparity but with filters of opposite phases, eliminates false matches between left and right visual features and yields a more accurate representation of disparity in the scene

(see also Read & Cumming, 2007). Furthermore, it can be proven theoretically that using divisive normalization to implement the aforementioned inhibitory interactions between phase disparities yields a representation that is the most accurate one that could possibly be achieved given the filter bank used (Kim, Mély & Serre, in preparation).

3.6. Intensity

We reused the brightness gradient from the $*Pb$ system, labeled "intensity" here. It is based on raw pixel intensities quantized to 32 levels with K-means in order to be consistent with other cues. Also note that due to the lack of a corresponding filter, the edge-based approach for this cue is not defined (see below).

3.7. Dimensionality reduction with PCA

Because the dimensionality of each channel varies across cues, we normalized the dimension of each channel output using principal component analysis (PCA) in order to make their comparison relevant. We kept the minimum number of dimensions (which varied across cues) after projection on the principal component vectors to account for 99% of the total variance.

3.8. Edges vs. textons

The output of any given cue corresponds to the response of a battery of oriented filters plus a center-surround channel followed by proper rectification and normalization. There are two major prescriptions for how to leverage these filter responses to detect object boundaries.

One approach is the *edge-based* approach, where the outputs of oriented filters are interpreted as reflecting the presence of a boundary. In this case, we attempted to classify whether a given position lies on a boundary, using the output of the cue as a feature vector. The key assumption is that the best-responding filter should line up with the boundary to be detected. However, this approach is often plagued by false positives, as many natural textures are characterized by strong oriented structures, which induces the detection of spurious edges.

The second approach is the *region-based* approach, which aims at partitioning an image using learned mid-level representations—fitting statistical models to various cues in each of a set of regions. This is the approach taken in the *Pb* system (Martin et al., 2004) and its derivatives *mPb* and *gPb* (Arbelaez et al., 2010), which yield state-of-the-art results on natural scenes.

The key operation in *Pb* is a χ^2 operator (also known as an oriented gradient operator). It first divides a local circular neighborhood in the image into two halves along some orientation. Then, it measures the difference between the empirical distributions of the cue values for each half, by taking a χ^2 histogram distance between binned estimates of these distributions. The distributions can be estimated from filter outputs that co-occured often throughout our dataset, corresponding to "universal textures" (Arbelaez et al., 2010) that are present in natural scenes. These can be identified by clustering filter bank outputs across scenes; the K centroids then correspond to universal texture prototypes, or *textons*. Each pixel is then assigned the integer index of the closest texton, and the χ^2 operator is applied to the resulting assignment map using K bins. We chose the same clustering algorithm as in Martin et al. (2004), K-means clustering, as well as the same number of cluster centers $K = 32$. In practice, we found that increasing the number K of textons seemed to have little effect on the overall diagnosticity of any model associated to each cue. We kept the same values as in Arbelaez et al. (2010) for the widths of the oriented gradient operator: 10, 20 and 40 pixels.

4. Results

4.1. Scale selection and diagnosticity of individual cues

For all the visual cues considered, we evaluated each scale and approach separately: for the edge-based approach, one classifier was considered for each filter size (trained on the raw filter outputs then evaluated against ground-truth). For the region-based approach, one classifier was considered for each filter size and χ^2 operator size (two scale parameters). Keeping all scales separate revealed discrepancies in accuracy between these as some scales capture local image structure better than others. As a result, we found that the best performing scale (for each cue and approach, either edge-based or region-based) performed on par with or better than a previous attempt to evaluate a cue by combining all scales.

Thus, considering the best parameters for each cue, we found color, luminance and intensity to be generally the most diagnostic cues, whereas motion and stereo appeared to be significantly less diagnostic in isolation (see Table 1). The full results, including the accuracy of each cue at non-optimal parameters, measured by different metrics such as the F-score (harmonic mean of precision and recall in a precision-recall paradigm), the precision (P), the recall (R) and the area under the precision-recall curve (AUC), can be found in Fig. S2.

We also compared the accuracies across conditions (region-based against edge-based, boundary annotations against edge annotations), filter sizes and scales for each cue in Fig. S3 (using individual classifier F-scores).

4.2. Edges vs. textons

A comparison between the best (*i.e.*, optimal parameters) edge-based and the best region-based approaches is shown in Table 1 for the boundary annotations. For each considered cue, the region-based approach was found to be more accurate than the edge-based approach. More surprising is the fact that the superiority of the region-based approach over the edge-based approach was not limited to the (higher-level) boundary annotations (corresponding to boundaries which belong to "things" or "parts of things") but generalized, for all except the single-opponent (SO) color cue, to (lower-level) edge-based annotations (corresponding to all edges and not only object boundaries; see Fig. 4).

This suggests that the region-based approach, despite being designed to capture higher-level boundaries in the image, remains a fairly local operator, as it does not seem to be overly penalized by the use of very low-level annotations. As expected, retrieving edge annotations was a much easier problem for either approach than higher-level boundary annotations although, as expected, the gap between the two approaches was reduced going from boundary annotations to edge annotations. In addition, a comparison of the accuracy of each approach on the two annotation sets suggested that the detection of (lower-level) edges remains a significantly easier task than retrieving more "semantic" boundaries.

4.3. Cue combination

To learn the optimal cue combination, we concatenated the outputs of the various classifiers trained on individual cues (one classifier per cue per filter size for the edge-based approach, and one classifier per cue per filter size and χ^2 size for the region-based approach) to form a vector that was then passed to a second-stage classifier. This second-stage classifier was trained to combine the confidence (posterior probability of boundary) of each individual classifier (63 classifiers total).

Table 1
Accuracy (F scores) of individual cues evaluated against boundary annotations, for the best (*i.e.*, optimal parameter values) edge-based and region-based classifiers. The region-based approach does consistently better.

Cue	Luminance	Color (SO)	Color (DO)
F, best edge-based	0.65 ± 0.01	0.61 ± 0.02	0.63 ± 0.02
F, best region-based	0.68 ± 0.02	0.65 ± 0.01	0.66 ± 0.02
Cue	Motion	Stereo	Intensity
F, best edge-based	0.61 ± 0.01	0.57 ± 0.01	N/A
F, best region-based	0.65 ± 0.02	0.58 ± 0.02	0.66 ± 0.01

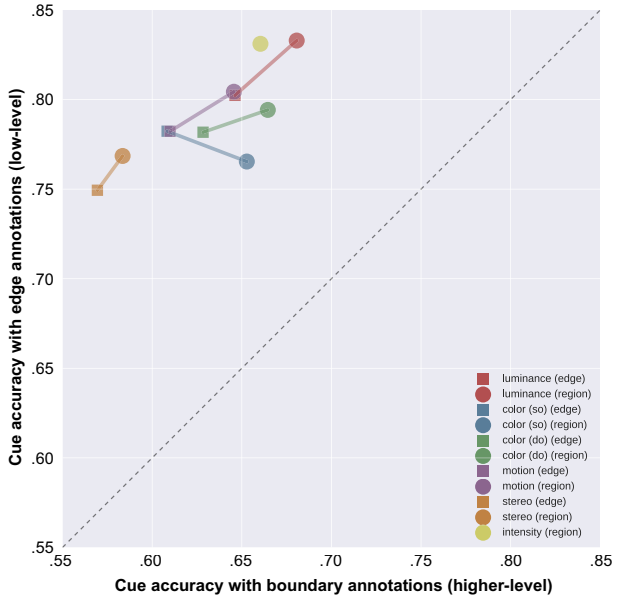

Fig. 4. Cue accuracy for the edge-based (squares) and the region-based (disks) approaches evaluated against both (high-level) boundary annotations (x-axis), and against (low-level) edge annotations (y-axis). All data-points are well above the unit diagonal, meaning the recovery of edge annotations is an easier challenge. Apart from color (SO), the region-based approach performs consistently better than the edge-based approach.

A summary of the accuracy measures obtained of the optimal cue combination as well as the three most individually diagnostic cues against both sets of annotations is given in Table 2. The learned cue combination was found to be more accurate that any cue in isolation (Fig. S2). Shown in Fig. 5 are true boundaries recovered by each cue for a sample frame. Shown in Fig. S4 are examples of classifier outputs for representative visual cues and their combination.

4.4. Cue selection and cumulative contribution to overall accuracy

To better understand how cues interact, we performed an analysis based on elastic net regularization (Zou & Hastie, 2005). This approach is similar to LASSO regularization but more robust to correlations between cues. In this framework, cue weights learned by the classifier incur a penalty proportional to the sum of their L^1 norm (causing most weights to be equal to zero, effectively resulting in cue selection) and L^2 norm. A regularization path is the value of a given cue's weight as a function of regularization strength. A high regularization strength forces most cues to have zero weight and, as the regularization strength is relaxed, the weights of the most diagnostic cues start growing and the classifier starts incorporating an increasing number of inputs/cues.

Hence, the earlier a cue is selected, the more important the cue is to the accuracy of the overall combination (see Fig. 6). Cues whose weights started to grow early, such as edge-based luminance with medium-sized filters, performed well in isolation, resulting in an early selection when strong regularization forced a small number of cues to participate in the combination.

Table 2
Accuracy (F score) of the optimal cue combination (Model) *vs.* ground-truth annotations (Human). Confidence intervals correspond to the s.e.m. Top cues are ranked according to their best F score across filter sizes, χ^2 sizes, and approaches (region *vs.* edge). Note that it is possible for computational models to exhibit a higher accuracy than the inter-annotator agreement, because a pixel marked as boundary is considered a true positive if it is labeled by any observer. The comparatively lower human level of agreement level is likely due to annotators' low individual recall given the overwhelmingly large number of admissible edges per scenes ($R = 0.65 \pm 0.041$).

	Human	Model	Top cues
Boundary annotations	0.76 ± 0.017	0.72 ± 0.014	1. Luminance 2. Color (DO) 3. Intensity
Edge annotations	0.75 ± 0.024	0.83 ± 0.004	1. Luminance 2. Intensity 3. Motion

We used the order of the selection by the elastic net regularized least square regression (which was stable across training/testing set splits) to establish a ranking of individual classifiers, which correspond to a cue (*e.g.*, luminance), an approach (either edge- or region-based), a filter size, and a χ^2 size (when applicable). Then, we progressively evaluated the accuracy of the combination (using the F score as before), starting from the top classifier, then adding the next best classifier, *etc.* until adding more did not make the combination more accurate (see Fig. 7 where each classifier was color-coded to indicate the cue it pertained to).

As expected from both the aforementioned classifier selection results and the individual classifier accuracies, intensity, luminance and color cues accounted for most of the combination's accuracy, with motion and stereo cues being amongst the last cues to contribute to the maximum observed accuracy. Only the top 28 classifiers, including all cues but prominently featuring luminance, intensity and color cues, were necessary to achieve the peak accuracy obtained when training all 63 available classifiers together. The first cues to be selected by the elastic net regression were quite correlated; for example, the correlation of the first 5 cues conditioned on boundary presence was over 0.50; the minimum correlation value found in this case was 0.32, and the maximum was 0.91 ($p < 0.01$ for all of these).

Interestingly, we found no clear relationship between the relative importance of individual cues for the overall combination (Fig. 7, lower x-axis), and their accuracy in isolation (Fig. 7, upper x-axis); *e.g.*, the fourth cue to have been selected was ranked 23rd out of 63 possible classifiers (see Fig. 7).

4.5. Why cue combination is effective

Fig. 8 (top) shows correlation matrices computed between the probabilities of a boundary being reported by each cue, conditioned on the presence or absence of a boundary in the human annotations. The detection probabilities were relatively uncorrelated close to boundary regions, thanks to an effective use of divisive normalization, and consistent with existing studies on the statistics of edges in natural scenes (Simoncelli & Schwartz, 1999; Zhou & Mel, 2008). It was still the case when considering the best classifier (*i.e.*, the best size for the filter and/or the χ^2 operator) for each cue, either edge-based or region-based (bottom). This also suggested another reason why combining cues is beneficial.

Moreover, between-cue correlations conditioned on the *absence* of a boundary (off-boundary conditional) in the annotations were higher than when conditioned on their presence (on-boundary

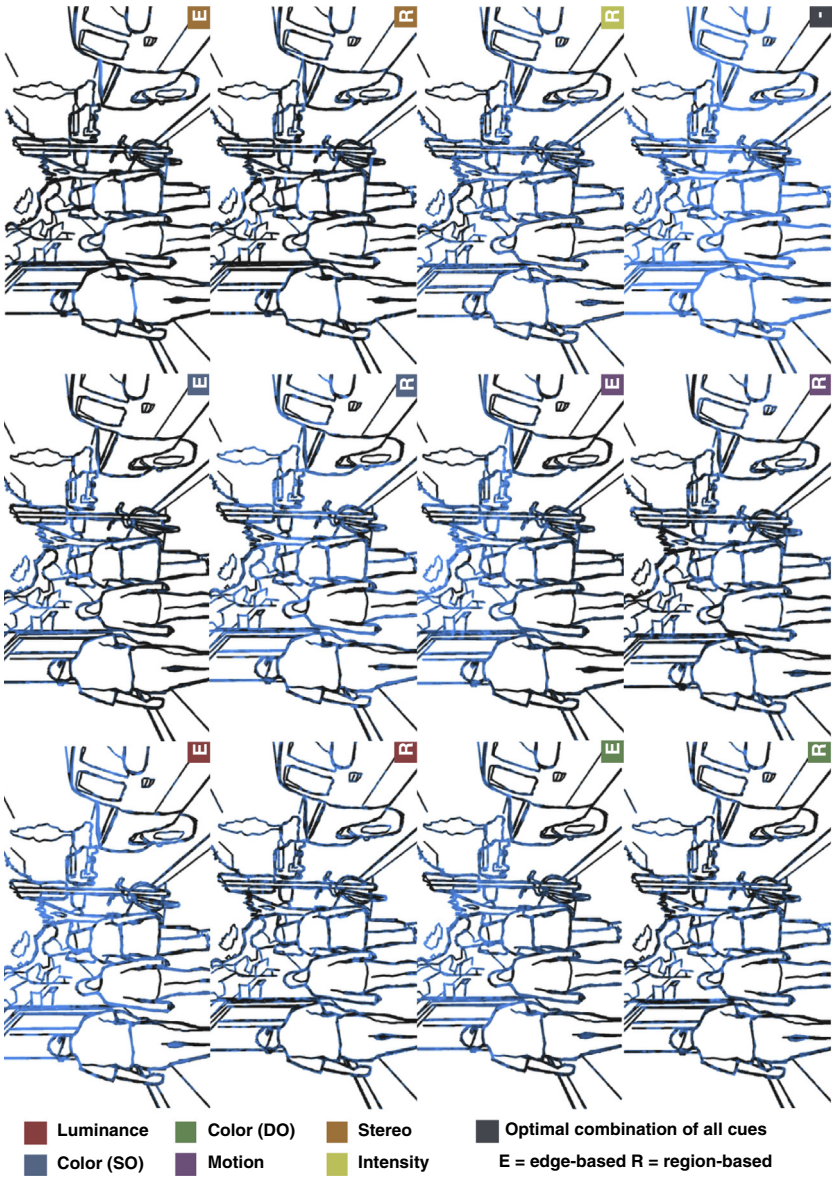

Fig. 5. Ground-truth boundaries (in black) for a representative scene shown in Fig. 1 (first column, second row). Parts of the ground-truth that were recovered by the corresponding cue are illustrated in blue, with the intensity of the blue color corresponding to the posterior probability of the boundary presence for that cue. This figure only shows recall; false alarms (boundary predictions from each cue not corresponding to a boundary in the ground-truth) are omitted for clarity.

conditional). This last result, which is robust even after excluding the weakest classifier responses, might seem surprising for classifiers designed to respond strongly to boundaries. One explanation is that they still respond to off-boundary regions weakly albeit reliably, thus driving up correlation. Note that these classifiers can still be considered fairly accurate under the precision-recall paradigm, where a high-enough threshold would ensure that spurious activity is not penalized as false positives. Similar observations can be made when considering the best performing classifiers for both edge-based and region-based approaches to each cue (see Fig. 8).

Fig. S4 shows examples of cue cooperation. For example, the right side of the person on the far left in the considered scene, which is clearly marked as a valid boundary in the ground-truth, is not at all detected by luminance or any other cue with the exception of the region-based approach for single-opponent color. Another example is the same person's right shoulder against the background, which receives roughly the same probability of being a boundary as background textures in luminance or double-opponent color. However, among other cues, stereo (region-based) correctly detects this contour to be a meaningful boundary, and does not respond against the aforementioned background. Hence, when all cues are combined, this boundary is assigned a higher probability of being a boundary than the neighboring background.

We also found that the classifiers trained on double-opponent color had consistent, vigorous responses to textured, off-boundary regions in natural scenes, in addition to picking up a subset of the annotated boundaries. As a result, most classifiers corresponding to double-opponent color cues had strong negative weights when trained in combination with all cues, despite the fact that double-opponent color was fairly diagnostic in isolation. Indeed, double-opponent color responded strongly to background texture edges, as well as object boundaries, at different levels of activity. When trained in isolation, it does well on its own as a boundary detector as long as an appropriate threshold is chosen

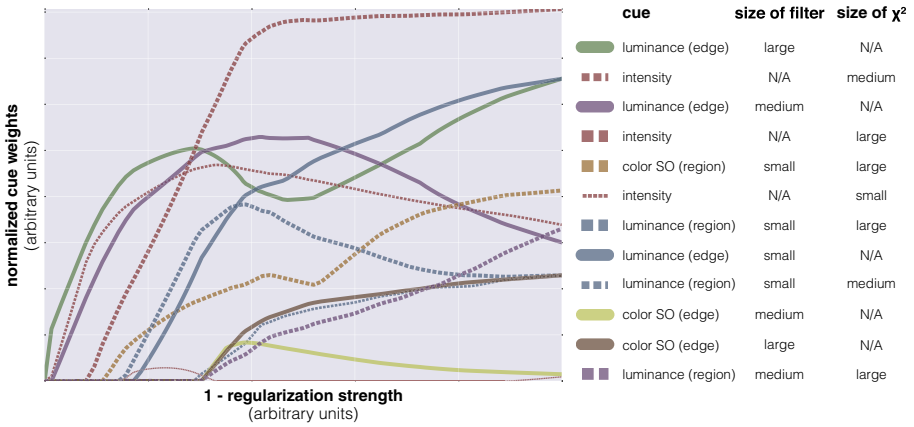

Fig. 6. Importance of individual cues for the combination measured using its corresponding regularization path. A high regularization strength forces most cues to have zero weight and, as the regularization strength progressively weakens, the weights associated with diagnostic cues start growing and additional cues get selected. Here we show the top 10 most diagnostic cues, which include cues that are selected simultaneously (which means they are strongly correlated).

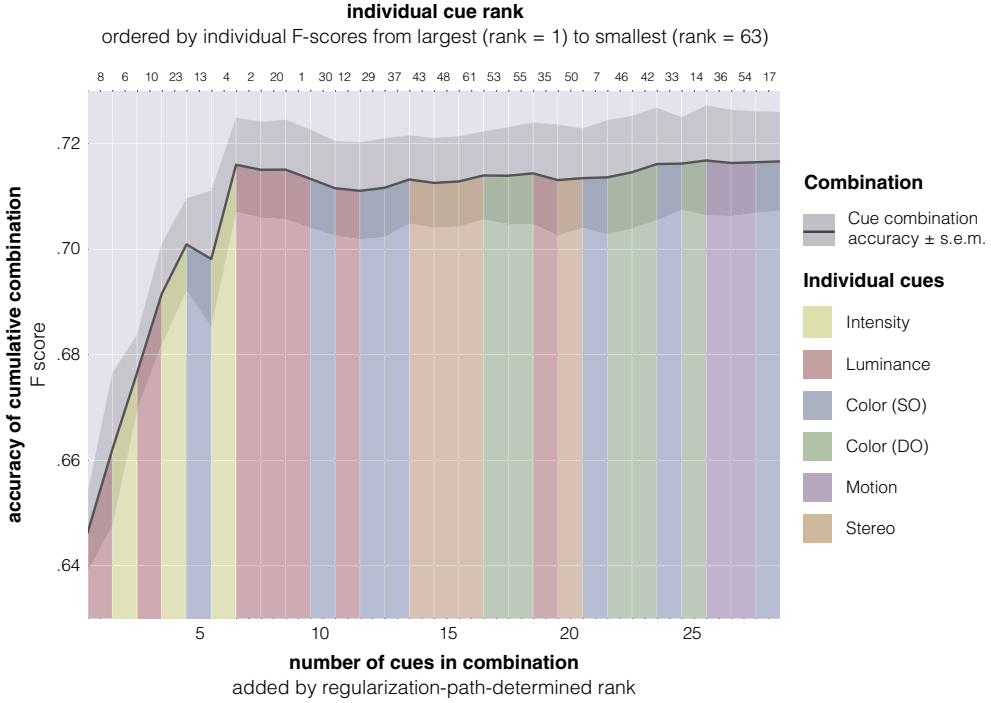

Fig. 7. Accuracy of the cue combination *vs.* number of cues added to the combination (lower *x*-axis). Cues were added given the ordering of the selection by the elastic net regularized least squares regression. The area under the accuracy curve is filled with colors indicating the cue being added to the combination (regardless of whether the approach is edge-based *vs.* region-based as well as the filter and χ^2 size). The upper *x*-axis shows cue ranks when evaluated in isolation (see Fig. S2 for the actual scores), between 1 (best cue, corresponding to luminance, region-based, filter size of "small" and χ^2 size of "medium") and 63 (not shown; worst cue, corresponding to stereo, region-based, filter size of "large" and χ^2 size of "small").

in the precision-recall framework: a threshold that is high enough so as to remove most of the background texture responses (thus helping precision), but low enough not to start removing actual boundaries and hurt its recall. However, in combination with other cues under a logistic model, subtracting the raw, non-thresholded output of double-opponent color (which is what a negative weight essentially does) from the rest of the cues lowers the probability that those background textures be interpreted as object boundaries. Effectively, this corresponds to a gating influence of double-opponent color during boundary detection, signaling spurious boundaries corresponding not to real boundaries but to textured regions with strong oriented content (see Fig. S4, where even the best classifier based on double-opponent color has a strong response to background textures).

5. Discussion

The aim of this study was to gain a deeper understanding of the image information available from individual early visual cues and their combination to support the detection of boundaries. We have thus implemented an integrated model of early visual processes in multiple visual cues from luminance to color, motion and binocular disparity. To assess the diagnosticity of these early visual cues, we collected color binocular video sequences of natural scenes to construct a video database with a set of higher-level object boundary annotations as well as lower-level edge annotations. We used the dataset and annotations to train and test machine-learning classifiers on these various visual channels for the detection of object boundaries in natural scenes.

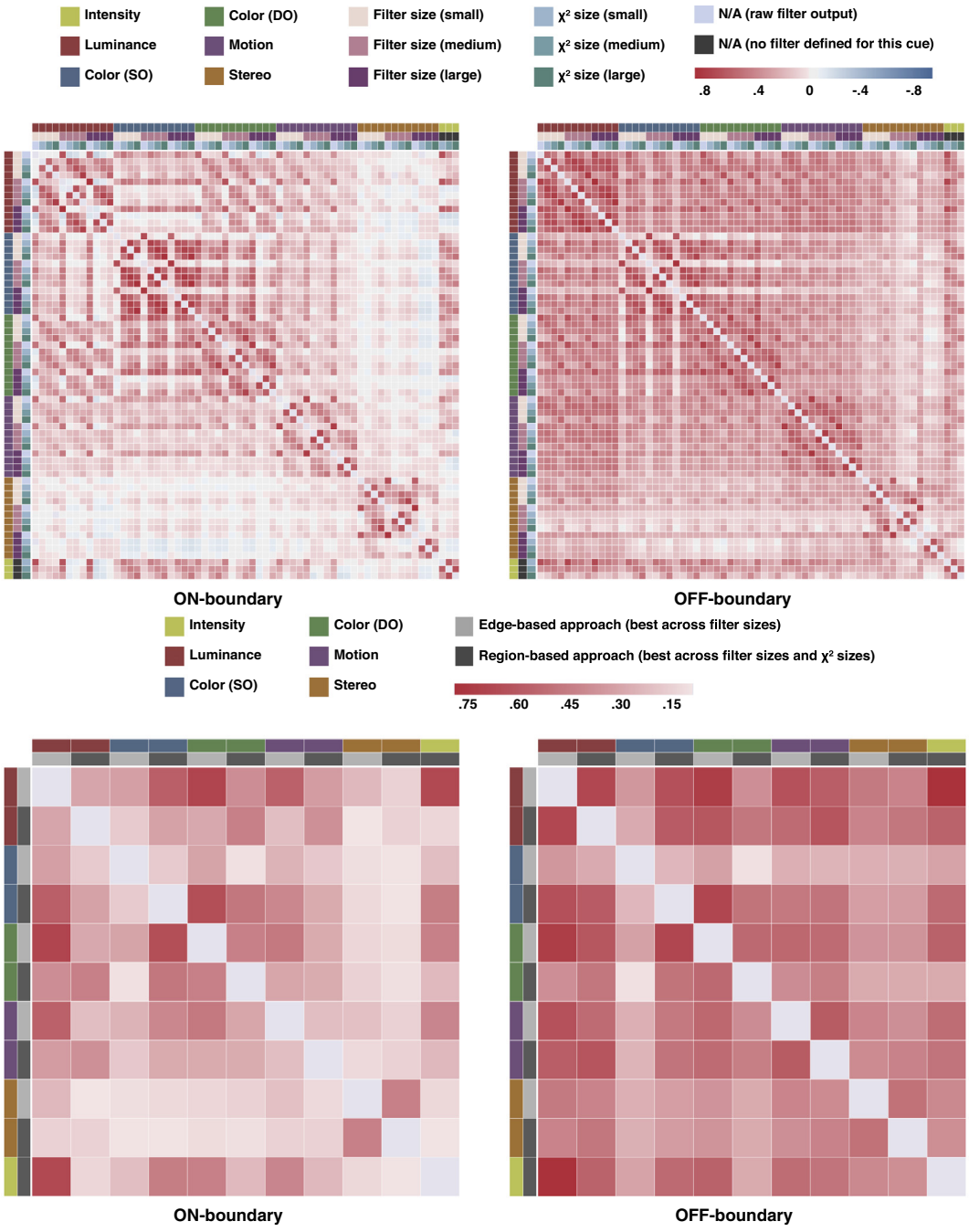

Fig. 8. Pearson's correlation between classifiers (probability of boundary presence) trained on individual cues, conditional on the presence (left) or absence (right) of a boundary in the ground-truths for each individual classifier trained on individual scales (top) and the best classifier for each cue across all scales. Each entry in the matrices corresponds to an average across all natural scenes, annotators, and training splits for each cue's classifiers. Note that most cues pick up edge-like structures in the input, whether they are part of a boundary or not. Thus, the rank order of the correlations looks very similar between ON- and OFF-boundary conditions.

In particular, we considered these early visual cues in the context of an edge-based approach where a population of orientation- and frequency-tuned filters selective for particular cues are used as direct inputs to a classifier trained to detect the presence or absence of a boundary. In a sense, the output of the filters is directly interpreted as reflecting the presence/absence of a boundary under the assumption that the best-responding filter should line up with the boundary to be detected. We also considered a region-based approach, which uses the population of orientation and frequency tuned filters to build a visual representation of intermediate complexity based on visual textons learned from images. In this second-order approach, boundaries are detected when two neighboring regions of the image are composed of different textures corresponding to different distributions of texton elements.

Our results confirmed that, for all cues considered, the region-based approach outperforms the edge-based approach for the detection of boundaries. Indeed, this result had already been shown in Martin et al. (2004) for the luminance cue and was expected because, in challenging natural scenes, textured object surfaces contain rich oriented structures which are likely to create numerous spurious detections when used as a direct measurement for the presence or absence of an edge.

Furthermore, a systematic comparison between the accuracies of individual cues suggested that luminance and color are much stronger cues than either motion or disparity. However, we also found that the optimal combination of all available cues yielded a classifier that fared significantly better than any cue in isolation. Previous work has shown that when luminance and chromatic contrasts are available, observers perform better than when either cues are presented in isolation (Frome et al., 1981; Rivest & Cavanagh, 1996). Our study thus adds to a growing body of literature providing a computational-level explanation for the observed improvement in accuracy when cues are combined.

In particular, we found an overall weak correlation between cues, especially at boundary locations. This result is consistent with an earlier study that looked at the statistics of luminance and chromatic edges in natural scenes and found chromatic information to be relatively non-redundant with luminance information (but see Fine et al., 2003; Zhou & Mel, 2008 for conflicting results). However, even though disparity and motion were among the least correlated with other cues, they only accounted for a (disappointingly) small increase in accuracy when included in combination with other cues. On the other hand, luminance, color, and intensity both accounted for a large portion of the accuracy in the optimal cue combination, and were selected early with respect to the regularization path analysis.

There are several limitations associated with our study. First, an inherent limitation in nearly every study of natural scene statistics is the absence of cue-specific ground-truth data. An alternative to this natural scene approach would be to rely on artificial stimuli that would enable the isolation of individual cues as in, for instance, moving random dot stimuli used in motion studies (Thompson, 1998). Here, instead, annotators were given ample time to scan images, zoom in and out, *etc.* under the assumption that, under these conditions, human annotators constitute a "golden standard" against which individual cues (motion, disparity, color, luminance) can be tested.

A major pitfall of our study is that the contribution of motion and stereo cues to overall boundary detection performance was disappointingly low, even when considering different metrics (individual F scores, importance in cue selection, *etc.*). Upon further inquiry, whereas, *e.g.*, color cues picked up on boundaries separating surfaces defined by different colors (*e.g.*, an animal's fur against a green background composed of leaves), motion cues were not as robust at picking up on boundaries between surfaces of different velocities, and binocular disparity cues on boundaries between surfaces placed at different depths. This comes as a surprise as there is an extensive literature suggesting that both motion and depth cues can be leveraged for boundary detection. But we have limited ourselves to early visual processing models of motion and stereo that include only one level of linear-nonlinear processing, resulting in populations of model V1 cells tuned to spatio-temporal frequency and absolute binocular disparity, respectively. Our results do not rule out motion and stereo as ineffective for boundary detection; but to successfully exploit them, models of higher-level processing may be required (such as an MT-level model of velocity tuning for motion, or a V2-level model of relative disparity tuning for stereo) that would possibly cascade several levels of linear-nonlinear processing.[2]

Indeed, it is known that individual V1 neuronal responses are ambiguous. For instance, direction-selective neurons suffer from the aperture problem (Bradley & Goyal, 2008) and disparity-tuned cells respond to false binocular matches (Parker, 2007) while neurons in higher level areas of the ventral and dorsal streams have largely solved these problems (Orban, 2008). Encouraging preliminary results were obtained with an extension of the motion processing model that is selective for velocity as opposed to spatio-temporal frequency only. This model groups surfaces moving at the same velocity and provided a small but significant improvement in accuracy. More generally, future work should assess the accuracy of hierarchical models of the visual cortex (see Mély & Serre, 2016 for a review) or their cousins, deep learning architectures (LeCun, Bengio, & Hinton, 2015), which incorporate multiple processing stages.

Another concern that limits the scope of the present study has to do with the way classifiers were trained. The benchmarking process that produced the F measure matches boundaries detected by the classifiers to those manually annotated while allowing for some slack in their precise localization (Martin et al., 2004) (to provide some robustness to small imprecisions in the labeling process). However, we did not allow for such slack during training because it would have yielded a prohibitively long training time. This might have penalized the resulting classifiers and it is possible that allowing for such slack during training would have yielded higher accuracies for both individual cues and their combinations. There is, however, no particular reason to expect that this would have affected some cues more than others and we thus expect that the resulting cue rankings would remain unchanged.

It is also possible that performance could be improved by considering more sophisticated classifiers and/or combination rules. The logistic classifier used here only affords a simple form of cooperation by linearly combining multiple outputs across cues. More sophisticated (non-linear) combination rules including random forests and other decision trees would theoretically be able to learn more complex interactions between cues.

Overall, we hope that the systematic assessment of the accuracy of early vision models for boundary detection, together with the resulting annotated video dataset, will provide a useful benchmark to gauge, progress in the development of higher-level vision models.

Acknowledgments

This work was supported by ONR grant (N000141110743), DARPA young faculty award (N66001-14-1-4037) and NSF early career award (IIS-1252951) to TS. Additional support was provided by the Center for Computation and Visualization (CCV) at Brown University.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.visres.2015.11.007.

References

Achanta, R., Shaji, A., Smith, K., Lucchi, A., Fua, P., & Süsstrunk, S. (2012). SLIC superpixels compared to state-of-the-art superpixel methods. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 34*(11), 2274–2282. http://dx.doi.org/10.1109/TPAMI.2012.120.

Adelson, E., & Bergen, J. (1991). The plenoptic function and the elements of early vision. *Computational Models of Visual Processing*, 3–20.

Adini, Y., Moses, Y., & Ullman, S. (1997). Face recognition: The problem of compensating for changes in illumination direction. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 19*(7), 721–732. http://dx.doi.org/10.1109/34.598229.

Albright, T. (1992). Form-cue invariant motion processing in primate visual cortex. *Science, 255*, 1141–1143 (80-.).

[2] From an engineering perspective, another venue to improve the performance of stereo and motion could be to ensure that the associated textons truly capture different depths and velocities, respectively, either by using a supervised method (*e.g.*, by decoding depth and speed) or keeping an unsupervised method like we did but also helping it by balancing the dataset with respect to the distribution of depths and speeds of the different surfaces represented in the dataset.

Arbelaez, P., Maire, M., Fowlkes, C., & Malik, J. (2010). Contour Detection and Hierarchical Image Segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1–20. http://dx.doi.org/10.1109/TPAMI.2010.161.

Black, M., & Fleet, D. (2000). Probabilistic detection and tracking of motion boundaries. *International Journal of Computer Vision, 38*(3), 231–245.

Bovik, A., Clark, M., & Geisler, W. (1990). Multichannel texture analysis using localized spatial filters. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 12*(1).

Bowyer, K., Kranenburg, C., & Dougherty, S. (2001). Edge detector evaluation using empirical ROC curves. *Computer Vision and Image Understanding, 84*(1), 77–103. http://dx.doi.org/10.1006/cviu.2001.0931.

Bradley, D., & Goyal, M. (2008). Velocity computation in the primate visual system. *Nature Reviews Neuroscience, 9*(9), 686–695. http://dx.doi.org/10.1038/nrn2472.

Caelli, T. (1993). Texture classification and segmentation algorithms in man and machines. *Spatial Vision, 7*(4), 277–292.

Carandini, M., & Heeger, D. (2012). Normalization as a canonical neural computation. *Nature Reviews Neuroscience, 13*, 51–62. http://dx.doi.org/10.1038/nrn3136.

Cavanagh, P. (1992). Multiple analyses of orientation in the visual system. In: *Front. Cogn. Neurosci.* (pp. 52–61).

Changizi, M. A., Zhang, Q., & Shimojo, S. (2006). Bare skin, blood and the evolution of primate colour vision. *Biology Letters, 2*(2), 217–221. http://dx.doi.org/10.1098/rsbl.2006.0440.

Chou, G. (1995). A model of figure-ground segregation from kinetic occlusion. In: *Proc. IEEE Int. Conf. Comput. Vis.* (pp. 1050–1057). doi: http://dx.doi.org/10.1109/ICCV.1995.466818.

Cremers, D., Rousson, M., & Deriche, R. (2006). A review of statistical approaches to level set segmentation: Integrating color, texture, motion and shape. *International Journal of Computer Vision, 72*(2), 195–215. http://dx.doi.org/10.1007/s11263-006-8711-1.

Derpanis, K., & Wildes, R. (2009). Early spatiotemporal grouping with a distributed oriented energy representation. In *IEEE conf. comput. vis. pattern recognit.* (pp. 232–239). IEEE. doi: http://dx.doi.org/10.1109/CVPR.2009.5206817.

Dollar, P., & Belongie, S. (2006). Supervised learning of edges and object boundaries. *2006 IEEE comput. soc. conf. comput. vis. pattern recognit.* (Vol. 2, pp. 1964–1971). IEEE. doi: http://dx.doi.org/10.1109/CVPR.2006.298.

Dominy, N. J., & Lucas, P. W. (2001). Ecological importance of trichromatic vision to primates. *Nature, 410*(6826), 363–366.

Ehinger, K. A., & Brockmole, J. R. (2008). The role of color in visual search in real-world scenes: Evidence from contextual cuing. *Perception & Psychophysics, 70*(7), 1366–1378.

Elder, J. H., & Goldberg, R. M. (2002). Ecological statistics of Gestalt laws for the perceptual organization of contours. *Journal of Vision, 2*(4), 324–353. doi: 10:1167/2.4.5.

Felzenszwalb, P. F., & Huttenlocher, D. P. (2004). Efficient graph-based image segmentation. *International Journal of Computer Vision, 59*(2), 167–181. http://dx.doi.org/10.1023/B:VISI.0000022288.19776.77.

Field, D. J., Hayes, A., & Hess, R. F. (1993). Contour integration by the human visual system: evidence for a local association field. *Vision Research, 33*(2), 173–193.

Fine, I., MacLeod, D. I. a., & Boynton, G. M. (2003). Surface segmentation based on the luminance and color statistics of natural scenes. *Journal of the Optical Society of America. A, Optics, Image Science, and Vision, 20*(7), 1283–1291.

Frome, F., Buck, S., & Boynton, R. (1981). Visibility of borders: separate and combined effects of color differences, luminance contrast, and luminance level. *Journal of the Optical Society of America, 71*(2), 145–150.

Geesaman, B. J., & Andersen, R. A. (1996). The analysis of complex motion patterns by form/cue invariant MSTd neurons. *Journal of Neuroscience, 16*(15), 4716–4732.

Gegenfurtner, K. R., & Rieger, J. (2000). Sensory and cognitive contributions of color to the recognition of natural scenes. *Current Biology, 10*(13), 805–808.

Geisler, W. S. (2008). Visual perception and the statistical properties of natural scenes. *Annual Review of Psychology, 59*, 167–192. http://dx.doi.org/10.1146/annurev.psych.58.110405.085632.

Geisler, W. S., Albrecht, D. G., Crane, a. M., & Stern, L. (2001). Motion direction signals in the primary visual cortex of cat and monkey. *Visual Neuroscience, 18*(4), 501–516.

Geisler, W. S., Perry, J. S., & Ing, A. D. (2008). Natural systems analysis. In: *SPIE Proc. 6806, Hum. Vis. Electron. Imaging, Vol. 6806*.

Geisler, W. S., & Perry, J. S. (2004). Contour statistics in natural images: Grouping across occlusions. *Visual Neuroscience, 26*(1), 109–121. http://dx.doi.org/10.1017/S0952523808080875.

Gelautz, M., & Markovic, D. (2004). Recognition of object contours from stereo images: An edge combination approach. In: *2nd Int. Symp. 3D Data Process. Vis. Transm* (pp. 774–780). doi: http://dx.doi.org/10.1109/TDPVT.2004.1335394.

Guo, Y., & Kimia, B. (2012). On evaluating methods for recovering image curve fragments. In *2012 IEEE comput. soc. conf. comput. vis. pattern recognit. work.* (pp. 9–16). IEEE.

Guo, Y., & Kimia, B. (2012). On evaluating methods for recovering image curve fragments. In *IEEE comput. soc. conf. comput. vis. pattern recognit. work.* (pp. 9–16). IEEE. doi: http://dx.doi.org/10.1109/CVPRW.2012.6238927.

Gupta, S., Arbel, P., & Malik, J. (2013). Perceptual organization and recognition of indoor scenes from rgb-d images. *IEEE Conference on Computer Vision and Pattern Recognition*, 564–571.

Hansen, T., & Gegenfurtner, K. R. (2009). Independence of color and luminance edges in natural scenes. *Visual Neuroscience, 26*(1), 35–49.

Hubel, D., & Wiesel, T. (1962). Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *Journal of Physiology, 160*, 106–154.

Humanski, R. A., & Wilson, H. R. (1993). Spatial-frequency adaptation: Evidence for a multiple-channel model of short-wavelength-sensitive-cone spatial vision. *Vision Research, 33*(5-6), 665–675.

Hurlbert, A. (1989). *The computation of color* (Ph.D. thesis).

Ing, A., Wilson, J., & Geisler, W. (2010). Region grouping in natural foliage scenes: Image statistics and human performance. *Journal of Vision, 10*, 1–19. http://dx.doi.org/10.1167/10.4.10.Introduction.

Jarrett, K., Kavukcuoglu, K., Ranzato, M. A., & LeCun, Y. (2009). What is the best multi-stage architecture for object recognition? In: *IEEE 12th Int. Conf. Comput. Vis.* (pp. 2146–2153). doi: http://dx.doi.org/10.1109/ICCV.2009.5459469.

Johnson, A. P., Kingdom, F. A., & Baker, C. L. Jr., (2005). Spatiochromatic statistics of natural scenes: First- and second-order information and their correlational structure. *Journal of the Optical Society of America A. Optics and Image Science, 22*(10), 2050. http://dx.doi.org/10.1364/JOSAA.22.002050.

Johnston, E. B., Cumming, B. G., & Landy, M. S. (1994). Integration of stereopsis and motion shape cues. *Vision Research, 34*(17), 2259–2275.

Kapadia, M. K., Ito, M., Gilbert, C. D., & Westheimer, G. (1995). Improvement in visual sensitivity by changes in local context: Parallel studies in human observers and in V1 of alert monkeys. *Neuron, 15*(4), 843–856.

Kaufmann-Hayoz, R., Kaufmann, F., & Stucki, M. (1986). Kinetic Contours in infants' visual perception. *Child Development, 57*(2), 292.

Koschan, A., & Abidi, M. (2005). Detection and classification of edges in color images. *Signal Processing Magazine, IEEE*, 64–73.

Krüger, N. (1998). Collinearity and parallelism are statistically significant second-order relations of complex cell responses. *Neural Processing Letters, 8*(2), 117–129.

Landy, M. S., & Kojima, H. (2001). Ideal cue combination for localizing texture-defined edges. *Journal of the Optical Society of America A, 18*(9), 2307. http://dx.doi.org/10.1364/JOSAA.18.002307.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature, 521*(7553), 436–444. http://dx.doi.org/10.1038/nature14539.

Leventhal, A. G., Thompson, K. G., Liu, D., Zhou, Y., & Ault, S. J. (1995). Concomitant sensitivity to orientation, direction, and color of cells in layers 2, 3, and 4 of monkey striate cortex. *Journal of Neuroscience, 15*(3), 1808–1818.

Leventhal, A., Wang, Y., Schmolesky, M., & Zhou, Y. (1998). Neural correlates of boundary perception. *Visual Neuroscience, 15*, 1107–1118.

Levinshtein, A., Stere, A., Kutulakos, K. N., Fleet, D. J., Dickinson, S. J., & Siddiqi, K. (2009). TurboPixels: Fast superpixels using geometric flows. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 31*(12), 2290–2297. http://dx.doi.org/10.1109/TPAMI.2009.96.

Li, Z. (1998). A neural model of contour integration in the primary visual cortex. *Neural Computation, 10*(4), 903–940.

Lim, J. J., Zitnick, C. L., & Dollar, P. (2013). Sketch tokens: A learned mid-level representation for contour and object detection. In: *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.* (pp. 3158–3165). doi: http://dx.doi.org/10.1109/CVPR.2013.406.

Lotto, R. B., Clarke, R., Corney, D., & Purves, D. (2011). Seeing in colour. *Optics & Laser Technology, 43*(2), 261–269.

Malik, J., & Perona, P. (1990). Preattentive texture discrimination with early vision mechanisms.

Martin, D. R., Fowlkes, C. C., & Malik, J. (2004). Learning to detect natural image boundaries using local brightness, color, and texture cues. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 26*(5), 530–549. http://dx.doi.org/10.1109/TPAMI.2004.1273918.

Mély, D., & Serre, T. (2016). Towards a system-level theory of computation in the visual cortex. In: *Comput. Cogn. Neurosci. Vis.*

Moore, A. P., Prince, S. J. D., Warrell, J., Mohammed, U., & Jones, G. (2008). Superpixel lattices. *2008 IEEE conf. comput. vis. pattern recognit.* (pp. 1–8). IEEE. doi: http://dx.doi.org/10.1109/CVPR.2008.4587471.

Mullen, K. T., & Losada, M. A. (1999). The spatial tuning of color and luminance peripheral vision measured with notch filtered noise masking. *Vision Research, 39*(4), 721–731.

Ohzawa, I. (1998). Mechanisms of stereoscopic vision: The disparity energy model. *Current Opinion in Neurobiology, 8*(4), 509–515.

Ohzawa, I., DeAngelis, G. C., & Freeman, R. D. (1990). Stereoscopic depth discrimination in the visual cortex: Neurons ideally suited as disparity detectors. doi: http://dx.doi.org/10.1126/science.2396096.

Orban, G. A. (2008). Higher order visual processing in macaque extrastriate cortex. *Physiological Reviews, 88*(1), 59–89. http://dx.doi.org/10.1152/physrev.00008.2007.

Parker, A. J. (2007). Binocular depth perception and the cerebral cortex. *Nature Reviews Neuroscience, 8*(5), 379–391.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research, 12*, 2825–2830.

Ramachandra, C. A., & Mel, B. W. (2013). Computing local edge probability in natural scenes from a population of oriented simple cells. *Journal of Vision, 13*(14). http://dx.doi.org/10.1167/13.14.19.

Read, J. C., & Cumming, B. G. (2007). Sensors for impossible stimuli may solve the stereo correspondence problem. *Nature Neuroscience, 10*(10), 1322–1328. http://dx.doi.org/10.1038/nn1951.

Regan, B. C., Julliot, C., Simmen, B., Viénot, F., Charles-Dominique, P., & Mollon, J. D. (1998). Frugivory and colour vision in Alouatta seniculus, a trichromatic platyrrhine monkey. *Vision Research, 38*(21), 3321–3327.

Ren, X., Bo, L., & Fox, D. (2012). RGB-(D) scene labeling: Features and algorithms. *IEEE Conference on Computer Vision and Pattern Recognition*, 2759–2766. http://dx.doi.org/10.1109/CVPR.2012.6247999.

Rivest, J., & Cavanagh, P. (1996). Localizing contours defined by more than one attribute. *Vision Research, 36*(1), 53–66.
Ross, J., Badcock, D. R., & Hayes, A. (2000). Coherent global motion in the absence of coherent velocity signals. *Current Biology, 10*(11), 679–682.
Sary, G., Vogels, R., & Orban, G. A. (1993). Cue-invariant shape selectivity of macaque inferior temporal neurons. *Science, 260*, 995–997 (80-.).
Segaert, K., Nygård, G. E., & Wagemans, J. (2009). Identification of everyday objects on the basis of kinetic contours. *Vision Research, 49*(4), 417–428.
Shapley, R., & Hawken, M. J. (2011). Color in the cortex: Single- and double-opponent cells. *Vision Research, 51*, 701–717. http://dx.doi.org/10.1016/j.visres.2011.02.012.
Shi, J., & Malik, J. (1998). Motion segmentation and tracking using normalized cuts. In: *IEEE Int. Conf. Comput. Vis.* (pp. 1154–1160).
Sigman, M., Cecchi, G. A., Gilbert, C. D., & Magnasco, M. O. (2001). On a common circle: Natural scenes and Gestalt rules. *Proceedings of the National Academy of Sciences of the USA, 98*(4), 1935–1940. http://dx.doi.org/10.1073/pnas.031571498.
Silberman, N., Hoiem, D., Kohli, P., & Fergus, R. (2012). Indoor segmentation and support inference from RGBD images. In: *Eur. Conf. Comput. Vis.* (pp. 1–14).
Simoncelli, E. P., & Heeger, D. J. (1998). A model of neuronal responses in visual area MT. *Vision Research, 38*(5), 743–761. doi:S0042698997001831 [pii].
Simoncelli, E. P., & Olshausen, B. A. (2001). Natural image statistics and neural representation. *Annual Review of Neuroscience, 24*, 193–216.
Simoncelli, E. P., & Schwartz, O. (1999). Modeling surround suppression in V1 neurons with a statistically-derived normalization model. *Advances in Neural Information Processing Systems*, 153–159.
Sincich, L., & Horton, J. (2005). The circuitry of V1 and V2: Integration of color, form, and motion. *Annual Review of Neuroscience, 28*(1), 303–326. http://dx.doi.org/10.1146/neuro.2005.28.issue-1.
Sundberg, P., Brox, T., Maire, M., Arbaelez, P., & Malik, J. (2011). Occlusion boundary detection and figure/ground assignment from optical flow. In *IEEE conf. comput. vis. pattern recognit.* (pp. 2233–2240). IEEE. doi: http://dx.doi.org/10.1109/CVPR.2011.5995364.
Tamrakar, A., & Kimia, B. B. (2007). No grouping left behind: from edges to curve fragments. In: *2007 IEEE 11th Int. Conf. Comput. Vis.* (pp. 1–8). doi: http://dx.doi.org/10.1109/ICCV.2007.4408919.
Tanabe, S., Haefner, R. M., & Cumming, B. G. (2011). Suppressive mechanisms in monkey V1 help to solve the stereo correspondence problem. *Journal of Neuroscience, 31*(22), 8295–8305. http://dx.doi.org/10.1523/JNEUROSCI.5000-10.2011.
Tassinari, H., Domini, F., & Caudek, C. (2008). The intrinsic constraint model for stereo-motion integration. *Perception, 37*(1), 79–95. http://dx.doi.org/10.1068/p5501.
Thompson, W. (1998). Exploiting discontinuities in optical flow. *International Journal of Computer Vision* (1985), 1–20.
Thurman, S., & Lu, H. (2013). Complex interactions between spatial, orientation, and motion cues for biological motion perception across visual space. *Journal of Vision, 13*, 1–18. http://dx.doi.org/10.1167/13.2.8.
Uttal, W. R., Spillmann, L., Stürzel, F., & Sekuler, A. B. (2000). Motion and shape in common fate. *Vision Research, 40*(3), 301–310.
VanRullen, R., Delorme, A., & Thorpe, S. (2001). Feed-forward contour integration in primary visual cortex based on asynchronous spike propagation. *Neurocomputing, 38-40*, 1003–1009. http://dx.doi.org/10.1016/S0925-2312(01)00445-3.
Veksler, O., Boykov, Y., & Mehrani, P. (2010). Superpixels and supervoxels in an energy optimization framework. In: *Eur. Conf. Comput. Vis.*
Voorhees, H., & Poggio, T. (1988). Computing texture boundaries from images. *Nature, 333*(6171), 364–367.
Vuong, Q. C., Hof, A. F., Bülthoff, H. H., & Thornton, I. M. (2006). An advantage for detecting dynamic targets in natural scenes. *Journal of Vision, 6*(1), 87–96. http://dx.doi.org/10.1167/6.1.8.
Wainwright, M. J., Schwartz, O., & Simoncelli, E. P. (2002). Natural image statistics and divisive normalization: Modeling nonlinearities and adaptation in cortical neurons. In: R. Rao, B. Olshausen, M. Lewicki (Eds.), *Stat. Theor. Brain*.
Wang, J. Y. A., & Adelson, E. H. (1994). Representing moving images with layers. *IEEE Transactions on Image Processing, 3*(5), 625–638.
Wermser, D., Liedtke, C. -E. (1982). Texture analysis using a model of the visual system. In: *Proc. Sixth Int. Conf. Pattern Recognit.* (pp. 1078–1080).
Woo, W., Kim, N., & Iwadate, Y. (2000). Object segmentation for Z-keying using stereo images. In: *5th Int. Conf. Signal Process. Proceedings, 2000. WCCC-ICSP 2000, Vol. 2* (pp. 1249–1254).
Wurm, L. H., Legge, G. E., Isenberg, L. M., & Luebker, A. (1993). Color improves object recognition in normal and low vision.
Young, M. J., Landy, M. S., & Maloney, L. T. (1993). A perturbation analysis of depth perception from combinations of texture and motion cues. *Vision Research, 33*(18), 2685–2696.
Zetzsche, C. (2001). Nonlinear mechanisms and higher-order statistics in biological vision and electronic image processing: Review and perspectives. *Journal of Electronic Imaging, 10*(1), 56. http://dx.doi.org/10.1117/1.1333056.
Zhang, J., Barhomi, Y., & Serre, T. (2012). A new biologically inspired color image descriptor. In: *Eur. Conf. Comput. Vis., Vol. 7576 LNCS* (pp. 312–324). doi: http://dx.doi.org/10.1007/978-3-642-33715-4_23.
Zhou, C., & Mel, B. W. (2008). Cue combination and color edge detection in natural scenes. *Journal of Vision, 8*(4), 4.1–4.2. http://dx.doi.org/10.1167/8.4.4.
Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic-net. *Journal of the Royal Statistical Society, 67*, 301–320.