Learning Affinity Functions for Image Segmentation: Combining Patch-based and Gradient-based Approaches

Abstract

This paper studies the problem of combining region and boundary cues for natural image segmentation. We employ a large database of manually segmented images in order to learn an optimal affinity function between pairs of pixels. These pairwise affinities can then be used to cluster the pixels into visually coherent groups. Region cues are computed as the similarity in brightness, color, and texture between image patches. Boundary cues are incorporated by looking for the presence of an "intervening contour", a large gradient along a straight line connecting two pixels.

We first use the dataset of human segmentations to individually optimize parameters of the patch and gradient features for brightness, color, and texture cues. We then quantitatively measure the power of different feature combinations by computing the precision and recall of classifiers trained using those features. The mutual information between the output of the classifiers and the same-segment indicator function provides an alternative evaluation technique which yields similar conclusions.

As expected, the best classifier makes use of brightness, color, and texture features, in both patch and gradient forms. We find that for brightness, the gradient cue outperforms the patch similarity. In contrast, using color patch similarity yields better results than using color gradients. Texture is the most powerful of the three channels, with both patches and gradients carrying significant independent information. Interestingly, the proximity of the two pixels does not add any information beyond that provided by the similarity cues. We also find that the convexity assumptions made by the intervening contour approach are supported by the ecological statistics of the dataset.

1. Introduction

Boundaries and regions are closely intertwined. A closed boundary generates a region while every image region has a boundary. Psychophysics experiments suggest that humans use both boundary and region cues to perform segmentation [43]. In order to build a vision system capable of parsing natural images into coherent units corresponding to surfaces and objects, it is clearly desirable to make global use of both boundary and region information.

Historically, researchers have focused separately on the sub-problems of boundary and region grouping. Region based approaches are motivated by the Gestalt notion of grouping by similarity. They typically involve integrating features such as color or texture over local patches of the image [8,12,32] and then comparing different patches [26,31]. However, smooth changes in texture or brightness caused by shading and perspective within regions pose a problem for this approach since two distant patches can be quite dissimilar despite belonging to the same image segment. To overcome these difficulties, gradient based approaches detect local edge fragments marked by sharp, localized changes in some image feature [4, 24, 18, 30, 20]. The fragments can then be linked together in order to identify extended contours [28, 42, 6].

Less work has dealt directly with the problem of finding an appropriate intermediate representation in order to incorporate non-closed boundary fragments into segmentation. Mathematical formulations outlined by [11, 25, 23] along with algorithms such as [19, 13] have attempted to utilize boundary and region information. More recently, [17, 40] have demonstrated the practical utility of integrating both in order to segment images of natural scenes.

There is a variety of widely told "folklore" regarding the various cues used for image segmentation: brightness gradients (caused by shading) and texture gradients (caused by perspective) necessitate a boundary-based approach; edge detectors are confused by texture, so one must use patch-based similarity for texture segmentation; color integrated over local patches is a robust and powerful cue. However, these contradictory statements have not been empirically challenged. By using a dataset of human segmentations [21, 5] as groundtruth, we are able to provide quantitative results regarding the ecological statistics[1] of patch- and gradient-based cues and gauge their relative effectiveness.

We treat the problem of integrating both gradient and patch information for segmentation within the framework of pairwise clustering [38,44,37,39,7,29,10,41]. In contrast to

[1] Our approach follows the lines of Egon Brunswik's suggestion nearly 50 years ago that the Gestalt factors made sense because they reflected the statistics of natural scenes [3].

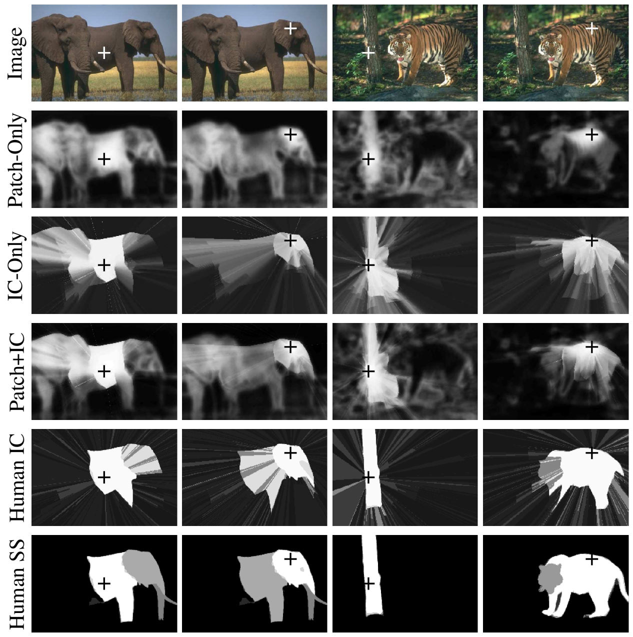


Figure 1: Pixel affinity images. The first row shows an image with one pixel selected. The remaining rows show the similarity between that pixel and all other pixels in the image, where white is most similar. Rows 2-4 show our patch-only, contour-only, and patch+contour affinity models. Rows 5 and 6 show the pixel similarity as given by the groundtruth data, where white corresponds to more agreement between humans. Row 6 shows simply the same-segment indicator function, while row 5 is computed using intervening contour on the human boundary maps.

central clustering techniques such as k-means or mixtures-of-Gaussians which compare each pixel (or other image element) to some small set of prototypes, pairwise techniques rely on the evaluation of an affinity function between each pair of image pixels. While pairwise techniques tend to be more computationally expensive, they have the advantage of removing the constraint that pixels be explicitly embedded in some normed vector space where Euclidean or Mahalanobis distances "make sense". Instead, pixels are implicitly described by their similarity to every other pixel in the image.

The pairwise framework allows patch and gradient information equal footing in the following way. Associate a descriptor to each pixel that captures color, brightness and texture in a neighborhood of the pixel. The patch based similarity between two pixels is a function of the difference in their descriptors. A gradient is computed as the change in these local descriptors between nearby pixels. For each pair of pixels, record the magnitude of the gradient encountered along a straight path connecting the two pixels in the image plane. Large gradients indicate the presence of an "intervening contour" [15] and suggests the pixels do not belong to the same segment. The pairwise affinity between the i-th and j-th pixel is given by a function whose arguments are the similarity between the i-th and j-th local descriptors and the gradients along the path from i and j.

Most applications of pairwise clustering to segmentation have made use of heuristically derived affinity functions (e.g. [17]). It is a natural proposal [22] to learn optimal pairwise affinities from training data. In the results presented here, nearly all free parameters (i.e. filter scales, histogram binning and quantization, descriptor windowing, combination of gradient features, etc.) have been carefully optimized with respect to training data. Our goal is to explicitly model the posterior probability of two pixels belonging to the same image segment conditioned on photometric properties of the image. Figure 1 shows examples of both groundtruth affinity functions and affinity models learned from data.

We provide two general schemes for evaluating the effectiveness of different combinations of features. The first is to train a classifier which declares two pixels as lying in the same or different segments given some set of features. Classifier performance is then evaluated by considering the trade-off between precision and recall. The second approach is to compute the mutual information between the classifier output and the same-segment indicator provided by the human segmentations. These two schemes are in strong agreement which lends force to our findings:

- Segmentations of the same image by different humans are quite consistent with each other. "Fine" segmentations tend to be "coarse" segmentations with regions that have been refined by breaking them into roughly convex parts.
- The ecological statistics of the dataset show that regions are mostly convex, validating the assumptions made by the intervening contour approach.
- Intervening contour and patch comparisons both provide significant, independent information about whether two pixels belong in the same segment.
- The color cue is best captured using patches, while for brightness one should use gradients. For texture, both gradients and patches are valuable.
- The proximity between two pixels does not provide any information not given by the patch-based or gradient-based similarity. It is simply a result of grouping, not a cue.

2. Methodology

We formulate the problem of learning the pixel affinity function as a classification problem of discriminating same-segment pixel pairs from different-segment pairs. Let S_{ij} be the true same-segment indicator so that S_{ij}=1 when pixels i and j are in the same segment, and S_{ij}=0 when pixels i and j are in different segments.

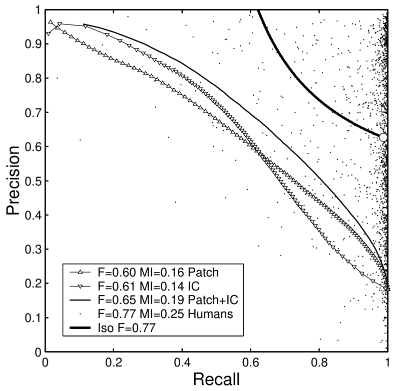


Figure 2: Performance of humans compared to our best pixel affinity models. The dots show the precision and recall of each of 1366 human segmentations in the 250-image test set when compared to the other humans' segmentation of the same image. The large dot marks the median recall (99%) and precision (63%) of the humans. The iso-F-measure curve at F=77% is extended from this point to represent the frontier of human performance for this task. The three remaining curves represent our patch-only model, contour-only model, and patch+contour model. Neither patches nor contours are sufficient, as there is significant independent information in the patch and contour cues. The model used throughout the paper is a logistic function with quadratic terms which performs the best among classifiers tried on this dataset.

The Berkeley Segmentation Dataset [21, 5] provides the groundtruth segmentation data. This dataset contains 12,000 manual segmentations of 1,000 images by 30 human subjects. Half of the images were presented to subjects in grayscale, and half in color. We use the color segmentations for 500 images, divided into test and training sets of 250 images each. Each image has been segmented by at least 5 subjects, so the groundtruth S_{ij} is defined by a set of human segmentations. We declare two pixels to lie in the same segment only if all subjects declare them to lie in the same segment.

Given a classifier output $\hat{S}_{ij}$, we can evaluate the classifier's performance in two ways. Our first evaluation technique uses the *precision-recall* (PR) framework, which is a standard method in the information retrieval community [33]. This framework was used by Abdou and Pratt [1] to evaluate edge detectors, and is similar to the ROC curve framework used by Bowyer et al. [2] for the same purpose. The approach produces a curve parameterized by detector threshold which shows the trade-off between noise and accuracy as the threshold varies. For example see Figure 2. *Precision* measures the probability that two pixels declared by the classifier to be in the same segment are in the same segment, i.e. $P(S_{ij}{=}1|\hat{S}_{ij}{=}1)$. *Recall* measures the probability that a same segment pair is detected, i.e. $P(\hat{S}_{ij} = 1|S_{ij} = 1)$. The PR approach is particularly appropriate when the two classes are unbalanced. By focusing on the scarcer class—same-segment pairs in our case—performance is not inflated by the ease of detecting the dominant class.

Precision and recall can be combined with the *F-Measure*, which is simply a weighted harmonic mean: $F = pr/(\alpha p + (1{-}\alpha)r)$. The weight α represents the relative importance of precision and recall for a particular application. We use $\alpha = 0.5$ in our experiments. The F-measure can be evaluated along the precision-recall curve, and the maximum value used to characterize the curve with a single number. When two precision-recall curves do not intersect, the F-measure is a useful summary statistic.

The second approach to evaluating a classifier measures the mutual information I between the classifier output $\hat{S}$ and the groundtruth data S. Given the joint distribution $p(x, y) = P(S{=}x, \hat{S}{=}y)$, the mutual information is defined as the Kullback-Liebler divergence between the joint and the product of the marginals, so $I(S; \hat{S}) = \int_{x,y} p(x, y) \log \frac{p(x,y)}{p(x)p(y)}$. We compute the joint distribution by binning the soft classifier output.

3. Features

We will model the affinity between two pixels as a function of both patch-based and gradient-based features. In each case, we can use brightness, color, or texture, producing a total of six features. We also consider the distance between the pixels as a seventh feature.

3.1. Patch-Based Features

Given a pair of pixels, we wish to measure the brightness, color, and texture similarity between circular neighborhoods of some radius centered at each pixel. Distributions of color in perceptual color spaces have been successfully used as region descriptors in image retrieval systems such as QBIC [26], as well as many color segmentation algorithms. We employ the 1976 CIE L*a*b* color space separated into luminance and chrominance features. We model brightness and color distributions with histograms constructed by binning a kernel density estimate of each feature. Histograms are compared with the χ^2 histogram difference operator [32].

For the brightness cue, we use the L* histogram for each pixel. In the case of color, it is not necessary to compute the joint a*b* histogram. Instead, it suffices to compute separate a* and b* histograms, and simply sum their χ^2 contributions. This is motivated by the fact that a* and b* correspond to the green-red and yellow-blue color opponent channels in the visual cortex, as well as the perceptual orthogonality of the two channels (see Palmer [27]).

The χ^2 histogram difference does not make use of the perceptual distance between the bin centers. Therefore, without smoothing, perceptually similar colors can have large χ^2 differences. Because the distance between nearby points in CIELAB space is perceptually meaningful inside a ball of some scale, binning a kernel density estimate whose kernel width σ matches this scale means that perceptually similar colors will have similar histogram contributions. Beyond this scale, where colors are incommensurate, χ^2 will regard them as equally different. The combination of a kernel density estimate in CIELAB with the χ^2 histogram difference is a good match to the structure of human color perception.

For the patch-based texture feature, we compare the distributions of filter responses in the two discs. There is an emerging consensus that for texture analysis, an image should first be convolved with a bank of filters tuned to various orientations and spatial frequencies [8, 18]. Our filter bank contains elongated quadrature pair filters—Gaussian second derivatives and their Hilbert transforms—at six orientations, along with one center-surround filter. The empirical distribution of filter responses has been shown to be a powerful feature for both texture synthesis [12] and texture discrimination [31].

There are many options for comparing the distributions (see Puzicha et al. [31]), but we use the approach developed in [17] which is based on the idea of *textons*. The texton approach estimates the joint distribution of filter responses using adaptive bins, which are computed with k-means. The texture descriptor for a pixel is therefore a k-bin histogram over the pixels in a disc of radius r centered on the pixel. As in [17], we compare descriptors with the χ^2 difference.

All of the patch-based features have parameters that require tuning, such as the radius of the discs, the binning parameters for brightness and color, and the texton parameters for texture. Section 4.3 covers the experiments that tune these parameters with respect to the training data.

3.2. Gradient-Based Features

Given a pair of pixels, consider the straight-line path connecting them in the image plane. If the pixels lie in different segments, then we expect to find, somewhere along the line, a photometric discontinuity or *intervening contour* [15]. If no such discontinuity is encountered, then the affinity between the pixels should be large.

In order to compute the intervening contour cue, we require a boundary detector that works robustly on natural images. For this we employ the gradient-based boundary detector of [20]. The output of the detector is a P_b image that provides the posterior probability of a boundary at each pixel. We consider the three P_b images computed using brightness, color, and texture gradients individually, as well as the P_b image that combines the three cues into a single boundary map. The combined model uses a logistic function trained on the dataset, which is well motivated by evidence in psychophysics that humans make use of multiple cues in localizing contours [34] perhaps using a linear combination [14]. Other classifiers besides the logistic function performed equally well.

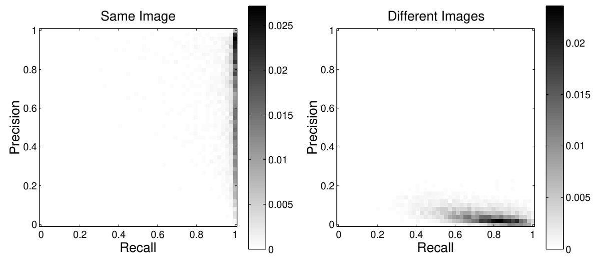


Figure 3: Agreement between human segmentations. The left panel shows the distribution of precision and recall for the 5555 human segmentations of all 1020 images in the dataset. The precision and recall are measured with respect to the class of same-segment pixel pairs, and each human is compared to the union of all other humans. High recall and lower precision supports the hypothesis that the different subjects perceive the same segmentation hierarchy, but segment at different levels of detail. The right panel shows the distribution of precision and recall when the left-out human and union-of-humans come from different images. This comparison provides a lower-bound for the similarity between segmentations without changing the statistics of the data.

The gradients are computed in a nearly identical manner to our patch features. Instead of comparing histograms between two whole discs, the gradient is based on the histogram difference between the two halves of a single disc, similar to [36, 35]. The orientation of the dividing diagonal sets the orientation of the gradient, and the radius of the disc sets the scale. All of the parameters of the gradients have been tuned by [20] on the same dataset to optimally detect the boundaries marked by the human subjects.

We compute the intervening contour cue for two pixels i and j from the P_b values that occur along the straight line path $\Gamma(t)$ connecting the two pixels. We consider the family of measures $L^p(\Gamma) = \left(\sum_t P_b(\Gamma(t))^p\right)^{1/p}$ for $p = \{0, 1, 2, 4, \infty\}$, as well as the mean of $P_b(\Gamma(t))$. The next section will cover the choice of the intervening contour function, as well as the best way to combine the contour information from the brightness, color, and texture channels.

4. Findings

4.1. Validating the Groundtruth Dataset

Before applying the human segmentation data to the problem of learning and evaluating affinity functions, we must determine that the dataset is self-consistent. To this end, we validate the dataset by comparing each segmentation to the remaining segmentations of the same image. Treating the left-out segmentation as the signal and the remaining segmentations as ground-truth, we apply our two evaluation methods.

The left panel of Figure 3 shows the distribution of precision and recall for the entire dataset of 1020 images. Since the "signal" in this case is binary-valued, we have a single point for each segmentation. The distribution is characterized by high recall, with a median value of 99%. This

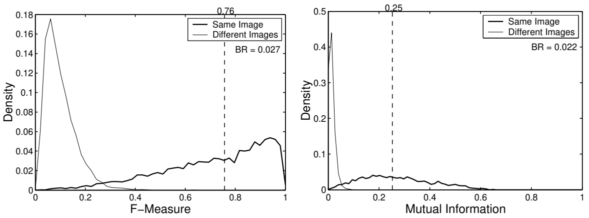


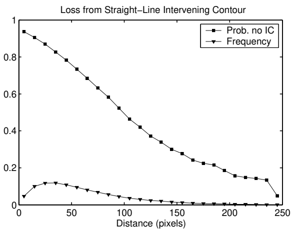
Figure 4: The left panel shows the same two distributions as Figure 3, with precision and recall combined using the F-measure. The right panel shows the distribution of mutual information for the same-image and different-image cases. The low overlap in each panel (2.7% and 2.2%) attests to the self-consistency of the data. As expected, significant information is shared between segmentations of the same image. The median F-measure of 0.76 and mutual information of 0.25 represents the target performance for our affinity models.

indicates that 99% of the same-segment pairs in the groundtruth are contained in a left-out human segmentation. The median precision is 66%, indicating that 66% of the same-segment pairs in a left-out human segmentation are contained in the ground-truth. These values are consistent with the interpretation that the different subjects provide consistent segmentations varying simply in their degree of detail. For comparison, the right panel of the figure shows the distribution when the left-out human segmentation and the groundtruth segmentations are of different images.

The left panel of Figure 4 shows the distribution of the F-measure for the same-image and different-image cases. Similarly, the right panel shows the distributions for mutual information. The clear separation between same-image and different-image comparisons attests to the consistency of segmentations of the same image. The median F-measure of 0.76 and mutual information of 0.25 represent the maximum achievable performance for pairwise affinity functions.

4.2. Validating Intervening Contour

Although the ecological statistics of natural images indicate that regions tend to be convex [9], the presence of an intervening contour does not necessarily indicate that two pixels belong in different segments. Concavities introduce intervening contours between same-segment pixel pairs. In this section, we analyze the frequency with which this happens.



Given the union of boundary maps for all human segmentations of an image, we measure the probability that same-segment pairs that have no intervening boundary contour. The figure at left shows this probability as a function of pixel separation, along with the number of same-segment pairs at each distance. If the regions were convex, then the curve would be fixed at one. Straight-line intervening contour is a good approximation to the same-segment indicator for small distances: 49% of pairs are in the $>$75% range.

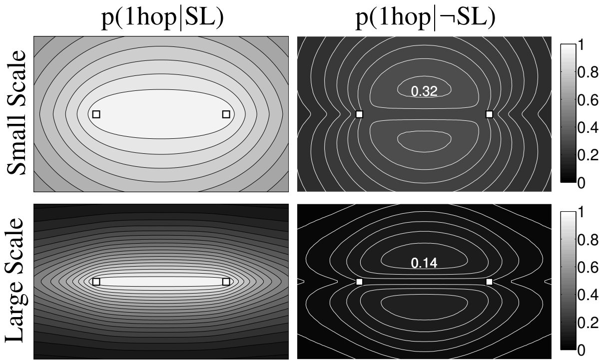


Figure 5: Each panel shows two pixels, marked with squares, that all humans declared to be in the same segment. The intensity at each point represents the empirical probability that a one-hop path through that point does not intersect a boundary contour. The left column is conditioned on there being an unobstructed straight-line (SL) path between the pixels, while the right column shows the probabilities when the SL path is obstructed. The top row shows data gathered from pixel pairs with small separation; the bottom row for pairs with large separation. See Section 4.2 for further discussion.

When straight-line intervening contour fails for a same-segment pair, there exist more complex paths connecting the pixels. Consider the set of paths that consist of two straight-line segments, which we call *one-hop* paths. The situations where one-hop paths succeed but straight-line paths fail can give us intuition about how much can be gained by examining paths more complex than straight line paths.

Figure 5 shows the empirical distribution of one-hop paths between same-segment pixel pairs, using the union of human segmentations. The probability of a one-hop path existing is conditioned on (left) there being a straight-line path with no intervening contour and (right) on there being no straight-line path. If the human subjects' regions were completely convex, then the right column images would be zero. Instead, we see that when straight-line intervening contour fails, there is a small but significant probability that a more complex one-hop path will succeed, and the probability of such a path is larger for smaller scales. There is clearly some benefit from the more complex paths due to concavities in the regions. However, the degree to which an algorithm could take advantage of the more powerful one-hop version of intervening contour depends on the frequency with which the one-hop paths find holes in the estimated boundary map. In any case, the figure makes clear that the simple straight-line path is a good first-order approximation to the connectivity of same-segment pairs.

Since the straight-line version of intervening contour will underestimate connectivity in concave regions, it may have a tendency toward over-segmentation. Figure 6 shows

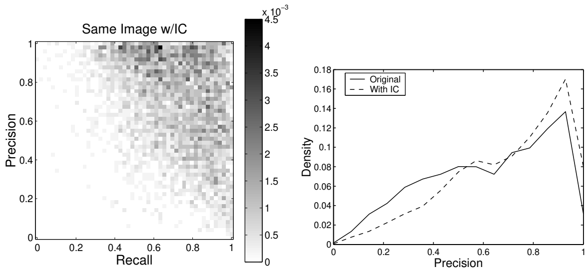


Figure 6: The potential over-segmentation caused by the intervening contour approach agrees with the refinement of objects by human observers. The distribution of precision and recall at left is generated in an identical manner as the left panel of Figure 3. However, we add the constraint that the same-segment pairs from the left-out human must not have an intervening boundary contour. The recall naturally decreases from adding a constraint to the "signal". However from the marginal distributions shown at right, we see that precision increases with the added constraint. Because the union segmentation is, on average, a refinement of the left-out segmentation, intervening contour tends to break non-convex regions in a manner similar to the human subjects.

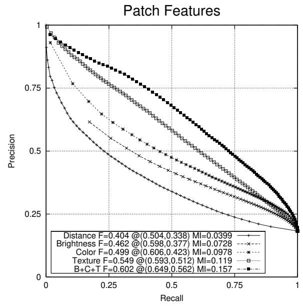
the effect on precision and recall for the human data when we add the constraint that same-segment pairs have no intervening boundary contour. As in Figure 3, we are comparing a left-out human to the union of the remaining humans. On average, the union segmentation will be more detailed than the left-out human. The figure shows a increase in median precision from 66% to 75%, indicating that intervening contour tends to break up non-convex segments in a manner similar to the human subjects. This lends confidence to an approach to perceptual organization of first finding convex object pieces through low-level processes, and then grouping the object pieces with into whole objects using higher-level cues.

4.3. Performance of Patches

Each of the brightness, color, and texture patch features has several parameters that require tuning. We optimized each patch feature independently via coordinate ascent to maximize performance on the training data. Figure 7 shows the result of the coordinate ascent experiments, where no change in any single parameter further improves performance.

For brightness and color, a radius of 5.76 pixels was optimal, though performance is similar for larger and smaller discs. In contrast, the texture disc radius has greater impact on performance, and the optimal radius is much larger at 16.1 pixels. The brightness and color patches also have parameters related to the binned kernel density estimates. The binning parameters for brightness are important for performance, while the color binning parameters are less critical. A larger σ indicates that small differences in the cue are less perceptually significant—or at least less useful for this task.

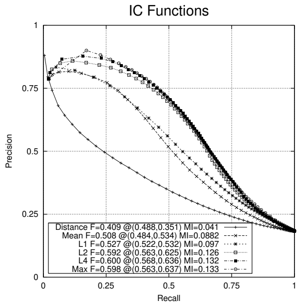
Apart from the disc radius, the texture patch cue has additional parameters related to the texton computation. The top right table in Figure 7 shows the optimization over the number of textons, with 512 being optimal. In general, we found that the number of textons should be approximately half the number of pixels in the disc. In addition, we find agreement with [20, 16] that the filter bank should contain a single scale, and that the scale should be as small as possible.

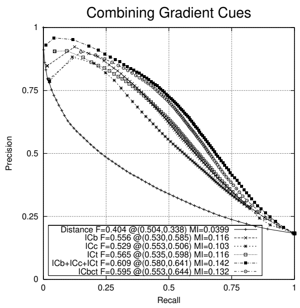


The graph at right shows the performance of classifiers trained on each patch feature individually, along with a classifier that uses all three. It is clear that each patch feature contains independent information, and that texture is the most powerful cue. The performance of a classifier that uses distance as its only cue is shown for comparison.

4.4. Performance of Gradients

The gradient cues are based on P_b images, which give the posterior probability of a boundary at each image location. The P_b function can incorporate any or all of the brightness, color, and texture cues, though consider for the moment the version that uses all three. Which intervening contour function should we use? The upper figure at left shows the performance of various functions including the mean, and the range of $L^p(\Gamma)$ functions from sum to max. The L^∞ version is clearly the best approach. Both the mean and the sum perform significantly worse. The results are the same no matter which cues the P_b function uses. Note that the max does not include any encoding of distance.





The lower figure at left compares the two ways in which we can combine the contour cues. We can either compute the intervening contour feature for brightness (ICb), color (ICc), and texture (ICt) separately and then combine with a classifier (ICb+ICc+ICt), or we can use the P_b function that combines the three channels into a single boundary map for the intervening contour feature (ICbct). We achieve better performance by computing separate contour cues.

Patch Radius

Radius	Brightness F	Brightness MI	Color F	Color MI	Texture F	Texture MI
0.010	0.46	.069	0.49	.093	-	-
0.014	0.46	.071	0.50	.096	-	-
0.020	**0.46**	**.074**	**0.50**	**.097**	-	-
0.028	0.46	.073	0.50	.097	0.50	.097
0.040	0.46	.071	0.49	.095	0.53	0.11
0.056	0.45	.067	0.48	.091	**0.55**	**0.12**
0.080	-	-	-	-	0.53	0.11
0.112	-	-	-	-	0.50	.089

Number of Textons

Num	F	MI
32	0.48	.081
64	0.50	.093
128	0.52	0.10
256	0.53	0.11
512	**0.55**	**0.12**
1024	0.52	0.10
2048	0.52	0.10

Kernel Density Estimate

Sigma	Bins	Brightness F	Brightness MI	Color F	Color MI
0.025	100	0.40	.041	0.48	.085
0.05	50	0.41	.047	0.50	.094
0.1	25	0.44	.059	**0.50**	**.097**
0.2	12	0.46	.070	0.491	.094
0.4	**6**	**0.46**	**.073**	0.49	.087
0.8	3	0.45	.062	0.48	.083

Texton Filter Bank

Scale	F	MI
0.007	**0.55**	**0.12**
0.010	0.55	0.11
0.014	0.53	0.11
0.020	0.51	.092
0.028	0.47	.072
0.007-0.014	0.55	0.12
0.010-0.020	0.53	0.11
0.014-0.028	0.51	.091

Figure 7: The parameters of the patches were optimized on the 250-image training set so that no change in any single parameter improved performance. The optimal patch sizes and filter scales are in units of the image diagonal, which is 288 pixels for our 240x160 images. The accessible ranges of the L*a*b* color axes were scaled to $[0, 1]$, which is the scale for the σ parameter. The Gaussian kernel was sampled at a constant rate of 21 samples per bin from $[-2\sigma, 2\sigma]$. We must reduce the number of bins as σ increases to keep the number of samples per bin constant. In the lower right table, the multi-scale texton filter bank contains three half-octave scales covering the range shown. See Section 4.3.

4.5. Cue Combination

We now have 7 prospective cues for our model of the pixel affinity, though we expect some to be redundant. The cues are brightness, color, and texture patches, intervening contour from the same three channels, and the distance between the two pixels in the image plane. We first evaluate the power of the distance cue in Figure 8. Whether we use a patch-only model, a contour-only model, or a patch+contour model, the result is always the same. Distance does not add any information not already provided by similarity cues.

We expect that the superiority of patch versus contour cues to differ depending on the feature channel. Smooth shading and foreshortening effects may favor brightness and texture gradients, while it is well known that color patches are a stable cue. Figure 9 shows the patch-only, contour-only, and patch+contour models for each of the brightness, color, and texture channels. As expected, the brightness patch proves to be far weaker than the brightness contour cue, with only marginal benefit from combining the two. Neither patches nor contours seem to dominate the color or texture channels. However, both texture cues appear quite powerful with independent information.

In order to determine the most fruitful combination of cues, we executed both top-down and bottom-up feature pruning experiments. Figure 10 shows the result. In both

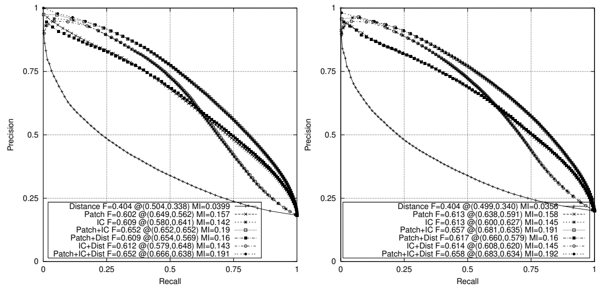


Figure 8: In this figure, we investigate the utility of the distance between two pixels as a cue for grouping. The Gestalt school identified proximity as a grouping cue, however, in all cases the classifier performance is the same whether or not distance is used. The right panel shows the same experiment with the test and training sets swapped. We performed all our experiments with swapped sets. Results were always consistent, with the F-measure and mutual information accurate to within two decimal places.

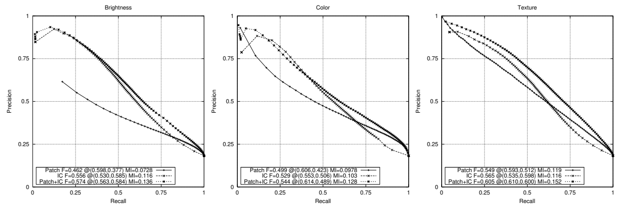


Figure 9: The three plots show classifiers that use either brightness (left), color (middle), or texture (right). Each plot shows the performance of a classifier using the patch cue, the gradient cue, and both together. The brightness patch appears an especially weak cue, which can be expected from the frequency of shading gradients in images. Both texture patches and texture gradients are powerful cues, and their combination is everywhere superior to using one alone.

cases, the model that maximizes performance using the fewest cues is the 4-cue model containing the brightness contour cue, the color patch cue, and both texture cues. All three feature channels are represented, with particular emphasis on texture. From the bottom-up pruning, it is clear that the texture cues are the most powerful along with color patches. It is interesting to see that at all stages in the pruning experiments, the model contains a balance between patch and contour cues, as well as a balance between the three channels.

4.6. Choice of Classifier

We find agreement with [20] that the choice of classifier is not important. Performance was always nearly identical whether we used a non-parametric density estimation method, or parametric models based on logistic regression, including simple logistic regression, logistic regression with quadratic features, or hierarchical mixtures of experts. To a first order approximation, a linear combination of features is sufficient. We favor the logistic with quadratic terms since it yields a slight improvement over the linear logistic function

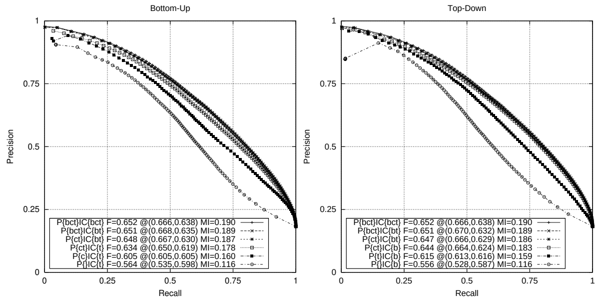


Figure 10: Feature pruning. In the left panel, we start with no features and add one feature at a time to the model in order to maximize performance in a greedy manner. In the right panel, we start with all six features, and greedily remove the worst feature, one feature at a time. In both cases, the model of choice uses the brightness contour, color patch, and both texture cues. The brightness patch and color contour are weak cues, while the color patch and both texture cues are powerful.

with little added computational cost.

5. Summary and Conclusions

We have shown how to combine patch and contour information into a model of pixel affinity for the purpose of image segmentation. For both patches and contours, we formulate brightness, color, and texture cues based on histogram differences. Contour cues are constructed in the intervening contour framework, which is justified by the ecological statistics of human segmentations. The six cues are carefully optimized with respect to a large dataset of manually segmented natural images, and then combined with a classifier trained on the groundtruth data. The modeled pixel affinity compares favorably to the human data using both precision/recall and mutual information measures.

References

[1] I. Abdou and W. Pratt. Quantitative design and evaluation of enhancement/thresholding edge detectors. *Proc. of the IEEE*, 67(5):753–763, May 1979.

[2] K. Bowyer, C. Kranenburg, and S. Dougherty. Edge detector evaluation using empirical ROC curves. 1999.

[3] E. Brunswik and J. Kamiya. Ecological validity of proximity and other Gestalt factors. *Am. J. Psych.*, pages 20–32, 1953.

[4] J. Canny. A computational approach to edge detection. *IEEE PAMI*, 8:679–698, 1986.

[5] Berkeley Segmentation Dataset, 2002. http://www.cs.berkeley.edu/projects/vision/bsds.

[6] J. Elder and S. Zucker. Computing contour closures. In *ECCV*, 1996.

[7] P. Felzenszwalb and D. Huttenlocher. Image segmentation using local variation. 1998.

[8] I. Fogel and D. Sagi. Gabor filters as texture discriminator. *Bio. Cybernetics*, 61:103–113, 1989.

[9] C. Fowlkes, D. Martin, and J. Malik. Understanding Gestalt cues and ecological statistics using a database of human segmented images. *POCV Workshop, ICCV*, 2001.

[10] Y. Gdalyahu, D. Weinshall, and M. Werman. Stochastic image segmentation by typical cuts. In *CVPR*, 1999.

[11] S. Geman and D. Geman. Stochastic relaxation, Gibbs distribution, and the Bayesian retoration of images. *IEEE PAMI*, 6:721–41, Nov. 1984.

[12] D. Heeger and J. Bergen. Pyramid-based texture analysis/synthesis. In *SIGGRAPH*, 1995.

[13] I. Jermyn and H. Ishikawa. Globally optimal regions and boundaries as minimum ratio weight cycles. *IEEE PAMI*, 23(10):1075–1088, 2001.

[14] M. Landy and H. Kojima. Ideal cue combination for localizing texture-defined edges. *J. Opt. Soc. Am. A*, 18(9):2307–2320, 2001.

[15] T. Leung and J. Malik. Contour continuity in region-based image segmentation. In *ECCV*, 1998.

[16] E. Levina. *Statistical Issues in Texture Analysis*. PhD thesis, University of California, Berkeley, 2002.

[17] J. Malik, S. Belongie, T. Leung, and J. Shi. Contour and texture analysis for image segmentation. *IJCV*, 43(1):7–27, June 2001.

[18] J. Malik and P. Perona. Preattentive texture discrimination with early vision mechanisms. *J. Opt. Soc. Am.*, 7(2):923–932, May 1990.

[19] R. Malladi, J. Sethian, and B. Vemuri. Shape modelling with front propogation: A level set approach. *IEEE PAMI*, 17(2):158–175, 1995.

[20] D. Martin, C. Fowlkes, and J. Malik. Learning to detect natural image boundaries using brightness and texture. 2002.

[21] D. Martin, C. Fowlkes, D. Tal, and J. Malik. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In *ICCV*, 2001.

[22] M. Meilă and J. Shi. Learning segmentation by random walks. In *NIPS*, 2001.

[23] Jean-Michel Morel and Sergio Solimini. *Variational Methods in Image Segmentation*. Birkhäuser, 1995.

[24] M. Morrone and D. Burr. Feature detection in human vision: a phase dependent energy model. *Proc. R. Soc. Lond. B*, 235:221–2245, 1988.

[25] D. Mumford and J. Shah. Optimal approximations by piecewise smooth functions, and associated variational problems. *Comm. in Pure and Applied Math.*, pages 577–684, 1989.

[26] W. Niblack et al. The QBIC project: Querying image by content using color, texture, and shape. *SPIE v. 1908*, 1993.

[27] S. Palmer. *Vision Science*. MIT Press, 1999.

[28] P. Parent and S. Zucker. Trace inference, curvature consistency, and curve detection. *IEEE PAMI*, 11(8):823–839, Aug. 1989.

[29] P. Perona and W. Freeman. A factorization approach to grouping. In *ECCV*, 1998.

[30] P. Perona and J. Malik. Detecting and localizing edges composed of steps, peaks and roofs. In *ICCV*, 1990.

[31] J. Puzicha, T. Hofmann, and J. Buhmann. Non-parametric similarity measures for unsupervised texture segmentation and image retrieval. 1997.

[32] J. Puzicha, Y. Rubner, C. Tomasi, and J. Buhmann. Empirical evaluation of dissimilarity measures for color and texture. In *ICCV*, 1999.

[33] C. Van Rijsbergen. *Information Retrieval, 2nd ed.* Dept. of Comp. Sci., Univ. of Glasgow, 1979.

[34] J. Rivest and P. Cavanagh. Localizing contours defined by more than one attribute. *Vision Research*, 36(1):53–66, 1996.

[35] Y. Rubner and C. Tomasi. Coalescing texture descriptors. *ARPA Image Understanding Workshop*, 1996.

[36] M. Ruzon and C. Tomasi. Color edge detection with the compass operator. In *CVPR*, 1999.

[37] S. Sarkar and K. Boyer. Quantitative measures of change based on feature organization: Eigenvalues and eigenvectors. In *CVPR*, 1996.

[38] G. Scott and H. Longuet-Higgins. Feature grouping by 'relocalisation' of eigenvectors of the proximity matrix. In *BMVC*, 1990.

[39] J. Shi and J. Malik. Normalized cuts and image segmentation. In *CVPR*, 1997.

[40] Z. Tu and S. Zhu. Image segmentation by data-driven markov chain monte carlo. *IEEE PAMI*, 24(5):657–673, May 2002.

[41] Y. Weiss. Segmentation using eigenvectors: a unifying view. *ICCV*, 1999.

[42] L. Williams and D. Jacobs. Stochastic completion fields: a neural model of illusory contour shape and salience. In *ICCV*, 1995.

[43] S. Wolfson and M. Landy. Examining edge- and region-based texture analysis mechanisms. *Vision Research*, 38(3):439–446, 1998.

[44] Z. Wu and R. Leahy. An optimal graph theoretic approach to data clustering: theory and its application to image segmentation. *IEEE PAMI*, 11:1101–13, Nov. 1993.