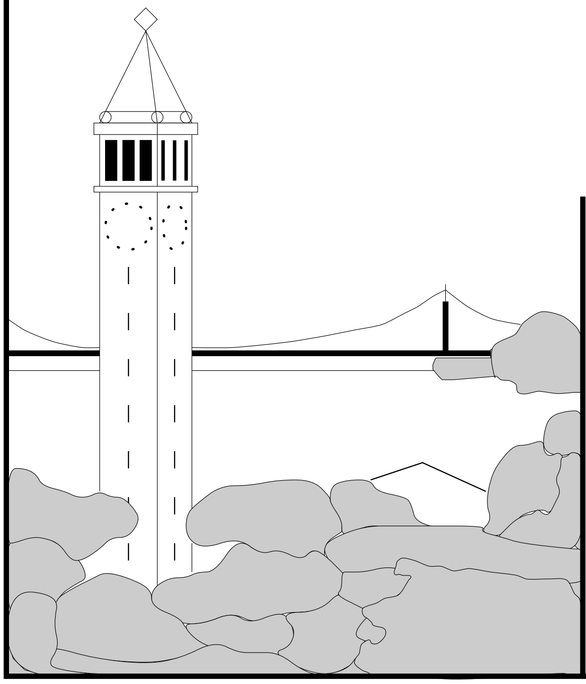
A Database of Human Segmented Natural Images and its Application to Evaluating Segmentation Algorithms and Measuring Ecological Statistics

David Martin, Charless Fowlkes, Doron Tal, Jitendra Malik



Report No. UCB/CSD-1-1133

January 2001

Computer Science Division (EECS)
University of California
Berkeley, California 94720

Abstract

This paper presents a database containing 'ground truth' segmentations produced by humans for images of a wide variety of natural scenes. We define an error measure which quantifies the consistency between segmentations of differing granularities and find that different human segmentations of the same image are highly consistent. Use of this dataset is demonstrated in two applications: (1) evaluating the performance of segmentation algorithms and (2) measuring probability distributions associated with Gestalt grouping factors as well as statistics of image region properties.

1. Introduction

Two central problems in vision are image segmentation and recognition[1]. Both problems are hard, and we do not yet have any general purpose solution approaching human level competence for either one.

While it is unreasonable to expect quick solutions to either problem, there is one dimension on which research in recognition is on much more solid grounds–it is considerably easier to quantify the performance of computer vision algorithms at recognition than at segmentation. Recognition is classification, and one can empirically estimate the probability of misclassification by simply counting classification errors on a test set. The ready availability of test sets – two of most significant ones are the MNIST handwritten digit dataset and the FERET face data set–has meant that different algorithms can be compared directly using the same quantitative error measures. It is well accepted that one cannot evaluate a recognition algorithm by showing a few images of correct classification. In contrast, image segmentation performance evaluation remains subjective. Typically, researchers will show their results on a few images and point out why the results 'look good'. We never know from such studies whether the results are best examples or typical examples, whether the technique will work only on images that have no texture, and so on.

The major challenge is that the question "What is a correct segmentation" is a subtler question than "Is this digit a 5". This has led researchers e.g. Borra and Sarkar[3]

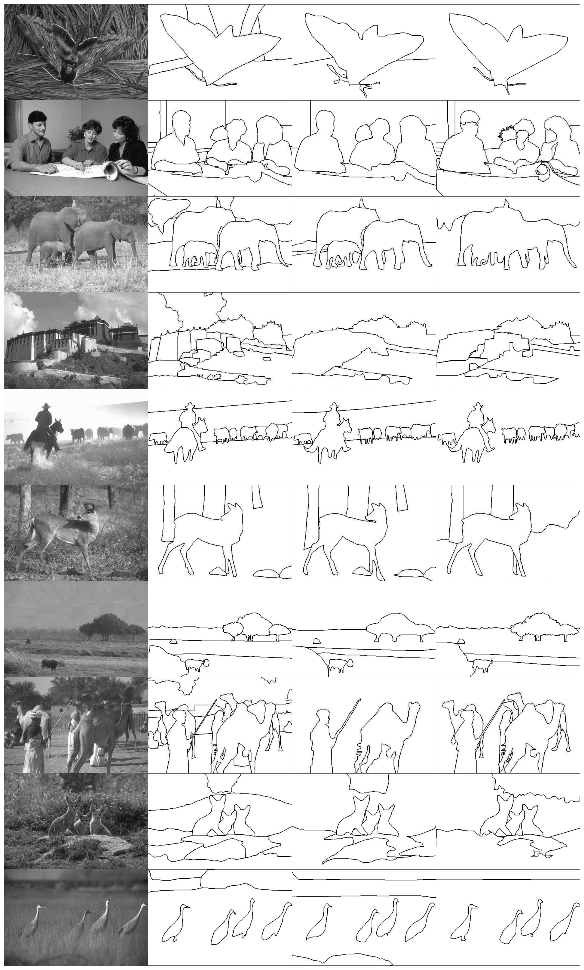


Figure 1: Sample of 10 images from the segmentation database. Each image has been segmented by 3 different people. A total of 10 people are represented in this data.

[1]It could be argued that they are aspects of the same problem. We do not necessarily disagree!

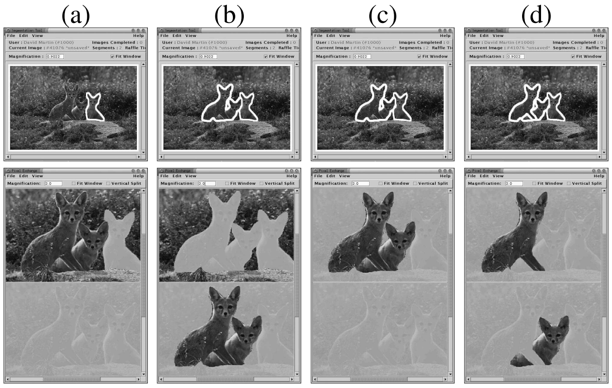


Figure 2: Using the segmentation tool. See §2.1 for details.

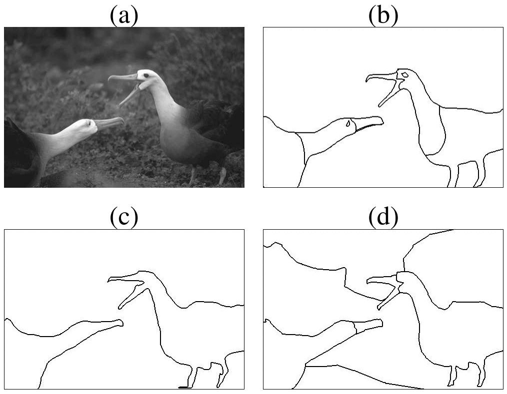


Figure 3: Motivation for making segmentation error measures tolerant to refinement. (a) shows the original image. (b)-(d) show three segmentations in our database by different subjects. (b) and (d) are both *simple refinements* of (c), while (b) and (d) illustrate *mutual refinement*.

to argue that segmentation or grouping performance can be evaluated *only* in the context of a task such as object recognition. We don't wish to deny the importance of evaluating segmentations in the context of a task. However, the thesis of this paper is that segmentations can also be evaluated purely as segmentations by comparing them to those produced by multiple human observers and that there is considerable consistency among different human segmentations of the same image so as to make such a comparison reliable.

Figure 1 shows some example images from the database and 3 different segmentations for each image. The images are of complex, natural scenes. In such images, multiple cues are available for segmentation by a human or a computer program–low level cues such as coherence of brightness, texture or continuity of contour, intermediate level cues such as symmetry and convexity, as well as high level cues based on recognition of familiar objects. The instructions to the human observers made no attempt to restrict or encourage the use of any particular type of cues. For instance, it is perfectly reasonable for observers to use their familiarity with faces to guide their segmentation of the image in the second row of Figure 1. We realize that this implies that a computational approach based purely on, say, low-level coherence of color and texture, would find it difficult to attain perfect performance. In our view, this is perfectly fine. We wish to define a 'gold standard' for segmentation *results* without any prior biases on what cues and algorithms are to be exploited to obtain those results. We expect that as segmentation and perceptual organization algorithms evolve to make richer use of multiple cues, their performance could continue to be evaluated on the same dataset.

Note that the segmentations produced by different humans for a given image in Figure 1 are not identical. But, are they consistent? One can think of a human's perceptual organization as imposing a hierarchical tree structure on the image. Even if two observers have exactly the same perceptual organization of an image, they may choose to segment at varying levels of granularity. See e.g. Figure 3. This implies that we need to define segmentation consistency measures that do not penalize such differences. We demonstrate empirically that human segmentations for the wide variety of images in the database are quite consistent according to these criteria, suggesting that we have a reliable standard with which to evaluate different computer algorithms for image segmentation. We exploit this fact to develop a quantitative performance measure for image segmentation algorithms.

There has been a limited amount of previous work evaluating segmentation performance using datasets with human observers providing the ground truth. Heath et al. [8] evaluated the output of different edge detectors on a subjective quantitative scale using the criterion of ease of recognizability of objects (for human observers) in the edge images. Closer to our work is the Sowerby image dataset that has been used by Huang [9] and Konishi et al. [12]. This dataset is small, not publicly available, and contains only one segmentation for each image. In spite of these limitations, the dataset has proved quite useful for work such as that of Konishi et al. who used it to evaluate the effectiveness of different edge filters as indicators of boundaries. We expect that our dataset would find far wider use, by virtue of being considerably more varied and extensive, and the fact that we provide a mechanism for computing the consistency of different segmentations.

The database that we have collected is a valuable resource for studying statistics of natural images. Most such studies in the past have concentrated on first and second order statistics such as the power spectrum or covariances, either on pixel brightnesses directly or on wavelet coefficients [10, 15, 16, 11, 12, 5, 13, 18]. We can go much further given the additional information provided by the segmentations. For instance, we can evaluate prior distributions corresponding to the various Gestalt factors such as similarity, proximity, convexity etc. and thus provide objective justifications for the use of these cues in grouping. While this way of thinking about the Gestalt factors was suggested nearly 50 years ago by Brunswik [4], so far empirical measurements of probability distributions have been limited to the factor of good continuation, e.g. [2]. Another application of the database is in studying the empirical distribution of sizes of regions in an image. This turns out to follow a power law, consistent with the work of Alvarez, Gousseau and Morel [1] with a rather different definition of sizes.

This paper is organized as follows. In § 2, we describe in detail the construction of the database of image segmentations. In § 3 we define measures for evaluating consistency of different segmentations of an image. § 4 puts the database to use by evaluating the performance of the Normalized cut algorithm on the different images. Performance is evaluated by computing the consistency of the computer segmentations with those made by human observers and comparing that to consistency among human observers. In § 5, we find another use for the database, namely in evaluating the ecological statistics of various Gestalt grouping factors. We conclude in § 6.

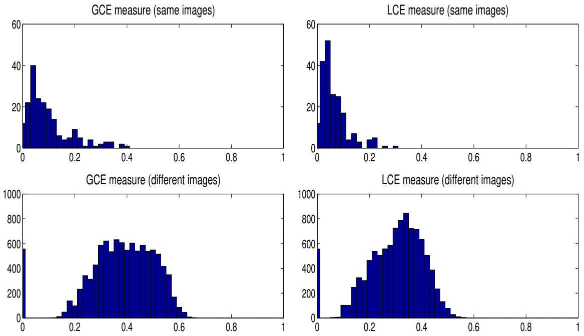


Figure 4: Distributions of the GCE (left) and LCE (right) measures over the segmentation database. The error measures are applied to all pairs of segmentations. The upper graphs show the error for segmentations of the same image. The lower graphs show the error for segmentations of different images. The spike at zero in the different-image graphs is due to degenerate segmentations of one particular image, of which everything else is a refinement. Clockwise from the top-left, the means are 0.11, 0.07, 0.39, 0.30.

2. Image Segmentation Database

The first task in constructing the segmentation database was to select a set of images. We chose 1000 representative 481x321 RGB images from the Corel image database. This database of 40,000 images is widely used in computer vision (e.g. [6, 7]). The criterion for selecting images was simple: We chose images of natural scenes that contain at least one discernible object. This criterion culls images that are inappropriate for the task of recognition, such as photographs of reflections of neon signs on wet concrete sidewalks, or photographs of marble textures.

2.1. Segmentation Tool

In order to easily collect segmentations from a wide range of people, we have developed a Java application that one can use to divide an image into segments, where a segment is simply a set of pixels. This approach has several advantages. First, anyone with Internet access can segment images. Second, the process produces an explicit partition of the pixels into groups (segments). Third, a server process can dynamically assign images to users, which gives precise control over the database content as it evolves.

Figure 2 shows a sequence of snapshots taken from a

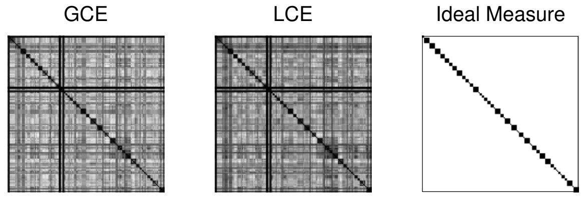


Figure 5: Error matrix for all image pairs, for GCE (left) and LCE (middle). M_{ij} corresponds to the error between segmentations i and j, where black signifies zero error. Segmentations are sorted by image, so segmentations of the same image are adjacent. The spurious horizontal and vertical bands confirm that the spike in the different-image graphs of Figure 4 are caused by degenerate segmentations of one image. The rightmost matrix shows the block-diagonal structure of the ideal error measure applied to a flawless dataset.

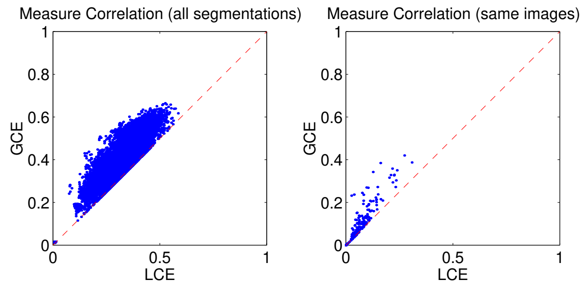


Figure 6: LCE vs. GCE for segmentations of different images (left) and the same image (right). The dashed line $x = y$ shows that GCE is a stricter measure than LCE.

typical session with the segmentation tool. Each snapshot shows two windows. The upper window is the main window of the application. It shows the image with all segments outlined in white. The lower window in each snapshot is the *splitter window*, which is used to split an existing segment into two new segments.

Consider Figure 2(a). The main window shows two segments. The user has selected the larger one in order to split it using the lower window. Between (a) and (b), the user drew a contour around the leftmost two pups in the top pane of the splitter window. This operation transfers the enclosed pixels to the bottom pane, creating a new segment. Between (c) and (d), the user split the two pups from each other. In (d), there are 4 segments.

In addition to simply splitting segments, the user can transfer pixels between any two existing segments. This provides a tremendous amount of flexibility in the way in which users create and define segments. The interface is simple, yet accommodates a wide range of segmentation styles. In less than 5 minutes, one can create a high-quality, pixel-accurate segmentation with 10-20 segments using a standard PC.

2.2. Experiment Setup and Protocol

It is imperative that variation among human segmentations of an image is due to different perceptual organizations of the scene, rather than aspects of the experimental setup. In order to minimize variation due to different interpretations of the task, the instructions were made intentionally vague in an effort to cause the subjects to break up the scene in a “natural” manner: *Divide each image into pieces, where each piece represents a distinguished thing in the image. It is important that all of the pieces have approximately equal importance. The number of things in each image is up to you. Something between 2 and 20 should be reasonable for any of our images.*

The initial subject group was a set of students in a graduate-level computer vision class who were additionally instructed to segment as naive observers. The subjects were provided with several example segmentations of simple, unambiguous images as a visual description of the task.

Images were assigned to subjects dynamically. When a subject requested a new image, an image was chosen randomly with a bias towards images that had been segmented by some other subject. In addition, the software ensured that (1) no subject saw the same image twice, (2) no image was segmented by more than 5 people, and (3) no two images were segmented by exactly the same set of subjects.

2.3. Database Status and Plans

The results in this paper were generated using our first version of the dataset that contains 150 grayscale segmentations by 10 people of 50 images, with 30 images with 3 or more segmentations. The data collection is ongoing, and at this time, we have 3000 segmentations by 25 people of 800 images. We aim to ultimately collect at least 4 grayscale and 4 color segmentations of 1000 images.

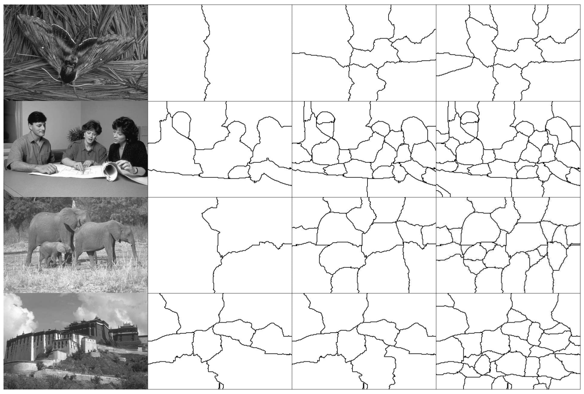


Figure 7: Segmentations produced by the Normalized Cuts algorithm using both contour and texture cues. Compare with Figure 1.

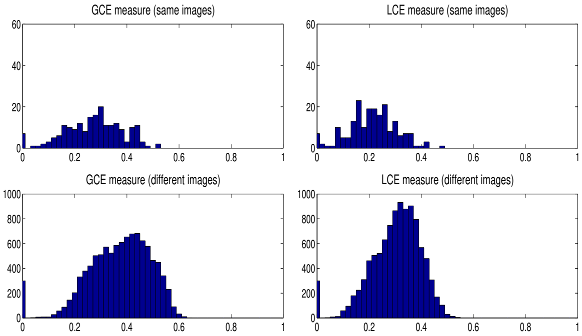


Figure 8: Distributions of the GCE (left) and LCE (right) measures for NCuts segmentations vs. human segmentations. The error measures were applied to pairs of segmentations, where each pair contains one NCuts and one human segmentations (see §4 for details). The upper graphs show the error for segmentations of the same image. For reference, the lower graphs show the error for segmentations of different images. Clockwise from the top-left, the means are 0.28, 0.22, 0.38, 0.31. Compare with Figure 4.

3. Segmentation Error Measures

There are two reasons to develop a measure that provides an empirical comparison between two segmentations of an image. First, we can use it to validate the segmentation database by showing that segmentations of the same image by different people are consistent. Second, we can use the measure to evaluate segmentation algorithms in an objective manner.

A potential problem for a measure of consistency between segmentations is that there is no unique segmentation of an image. For example, two people may segment an image differently because either (1) they perceive the scene differently, or (2) they segment at different granularities. If two different segmentations arise from different perceptual organizations of the scene, then it is fair to declare the segmentations inconsistent. If, however, one segmentation is simply a refinement of the other, then the error should be small, or even zero. Figure 3 shows examples of both simple and mutual refinement from our database. We do not penalize simple refinement in our measures, since it does not preclude identical perceptual organizations of the scene.

In addition to being tolerant to refinement, any error measure should also be (1) independent of the coarseness of pixelation, (2) robust to noise along region boundaries, and (3) tolerant of different segment counts between the two segmentations. The third point is due to the complexity of the images: We need to be able to compare two segmentations when they have different numbers of segments. In the remainder of this section, we present two error measures that meet all of the aforementioned criteria. We then apply the measures to the database of human segmentations.

3.1. Error Measure Definitions

A segmentation is simply a division of the pixels of an image into sets. A segmentation error measure takes two segmentations S_1 and S_2 as input, and produces a real-valued output in the range $[0..1]$ where zero signifies no error.

We define a measure of error at each pixel that is tolerant to refinement as the basis of both measures. For a given pixel p_i consider the segments in S_1 and S_2 that contain that pixel. The segments are sets of pixels. If one segment is a proper subset of the other, then the pixel lies in an area of refinement, and the local error should be zero. If there is no subset relationship, then the two regions overlap in an inconsistent manner. In this case, the local error should be non-zero. Let $\setminus$ denote set difference, and $|x|$ the cardinality of set x. If $R(S, p_i)$ is the set of pixels corresponding to the region in segmentation S that contains pixel p_i, the local refinement error is defined as:

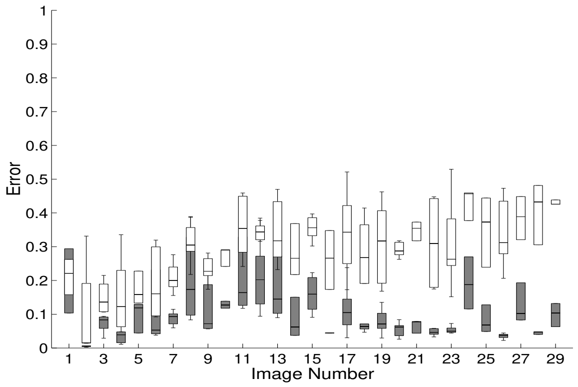


Figure 9: The GCE for human vs. human (gray) and NCuts vs. human (white) for each image for which we have ≥ 3 human segmentations. The LCE data is similar.

$$E(S_1, S_2, p_i) = \frac{|R(S_1, p_i) \backslash R(S_2, p_i)|}{|R(S_1, p_i)|} \qquad (1)$$

Note that this local error measure is not symmetric. It encodes a measure of refinement in one direction only: $E(S_1, S_2, p_i)$ is zero precisely when S_1 is a refinement of S_2 at pixel p_i, but not vice versa. Given this local refinement error in each direction at each pixel, there are two natural ways to combine the values into a error measure for the entire image. Global Consistency Error (GCE) forces all local refinements to be in the same direction. Local Consistency Error (LCE) allows refinement in different directions in different parts of the image. Let n be the number of pixels:

$$GCE(S_1, S_2) = \frac{1}{n} \min \left\{ \sum_i E(S_1, S_2, p_i), \sum_i E(S_2, S_1, p_i) \right\} \qquad (2)$$

$$LCE(S_1, S_2) = \frac{1}{n} \sum_i \min \left\{ E(S_1, S_2, p_i), E(S_2, S_1, p_i) \right\} \qquad (3)$$

As $LCE \leq GCE$ for any two segmentations, it is clear that GCE is a tougher measure than LCE. Looking at Figure 3, GCE would tolerate the simple refinement from (c) to (b) or (d), while LCE would also tolerate the mutual refinement of (b) and (d). Note that since both measures are tolerant of refinement, they are meaningful only when comparing two segmentations with an approximately equal number of segments. This is because there are two trivial segmentations that achieve zero error: One pixel per segment, and one segment for the entire image. The former is a refinement of any segmentation, and any segmentation is a refinement of the latter.

3.2. Error Measure Validation

We apply the GCE and LCE measures to all pairs of segmentations in our dataset with two goals. First, we hope to show that given the arguably ambiguous task of segmenting an image into an unspecified number of segments, different people produce consistent results on each image. Second, we hope to validate the measures by showing that the error between segmentations of the same image is low, while the error between segmentations of different images is high.

Figure 4 shows the distribution of error between pairs of human segmentations. The top graphs show the error between segmentations of the same image; the bottom graphs show the error between segmentations of different images. As expected, the error distribution for segmentations of the same image shows a strong spike near zero, while the error distribution for segmentations of different images is neither localized nor close to zero.

We characterize the separation of the two distributions by noting that for LCE, 5.9% of segmentation pairs lie above 0.12 for the same image or below 0.12 for different images. For GCE, 5.9% of pairs lie above 0.16 for the same image or below 0.16 for different images. Note the good behavior of both measures despite the fact that the number of segments in each segmentation of a particular image can vary by a factor of 10. Figure 5 shows the raw data used to the compute the histograms.

In Figure 6, we plot LCE vs. GCE for each pair of segmentations. As expected, we see (1) that GCE and LCE are measuring similar qualities, and (2) that GCE > LCE in all cases.

4. A Segmentation Benchmark

In this section, we use the segmentation database and error measures to evaluate the Normalized Cuts (NCuts) image segmentation algorithm.

In collecting our dataset, we permitted a great deal of flexibility in how many segments each subject created for an image. This is desirable from the point of view of creating an information-rich dataset. However, when comparing a human segmentation to a computer segmentation, our measures are most meaningful when the number of segments is approximately equal. For example, an algorithm could thwart the benchmark by producing one segment for the whole image, or one segment for each pixel. Due to the tolerance of GCE and LCE to refinement, both of these degenerate segmentations have zero error.

Since image segmentation is an ill-posed problem without stating the desired granularity, we can expect any segmentation algorithm to provide some sort of control over the number of segments it produces. If our human segmentations of an image contain 4, 9, and 13 segments, then we instruct the computer algorithm to also produce segmentations with 4, 9, and 13 segments. We then compare each computer segmentation to each human segmentation. In this way, we can make a meaningful comparison to the human segmentation error shown in Figure 4. In addition, we consider the mean error over all images as a summary statistic that can be used to rank different segmentation algorithms.

The NCuts algorithm [17, 14] takes a graph theoretic approach to the problem of image segmentation. An image is treated as a weighted graph. Each pixel corresponds to a node, and edge weights computed from both contour and texture cues denote a local measure of similarity between two pixels. NCuts segments an image by cutting this graph into strongly connected parts. The version of NCuts described in [14] automatically determines the number of regions by splitting the graph until the cuts surpass a threshold. We modified the stopping criterion to provide explicit control over the final number of segments.

Figure 8 shows the error between NCuts segmentations and human segmentations. In comparing this NCuts error to the human error shown in Figure 4, we see that NCuts is producing segmentations worse than humans, but still better than "random." The error distributions for segmentations of different images (the bottom graphs in each figure) approximate the performance of random segmentation. The mean error over all segmentation pairs gives NCuts an overall error of 22% by LCE (compared to 7% for humans), and 28% by GCE (compared to 11% for humans).

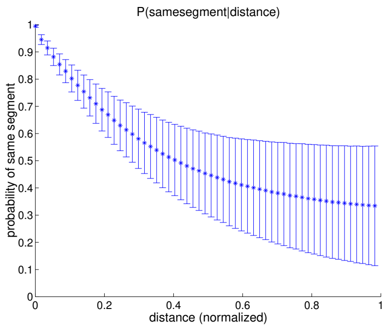


Figure 10: Proximity: The probability that two points belong to the same segment given their distance. Distances have been scaled per image as discussed in the text and normalized to range from 0 to 1. We sample 1000 points from each segmentation and compute all pairwise distances. Error bars show $\pm\sigma$ intervals.

Figure 9 shows both the human error (blue) and NCuts error (red) for each image separately. In most cases, the human segmentations form a tight distribution near zero. In virtually all cases, NCuts performs worse than humans, but it fares better on some images than others. This data can be used to find the type of images for which an algorithm has the most difficulty.

5. Bayesian Interpretation of Gestalt Grouping Factors

Brunswik [4] suggested that the various Gestalt factors of grouping such as proximity, similarity, convexity, etc. made sense because they reflected the statistics of natural scenes. For instance, if nearby pixels are more likely to belong to the same region, it is justified to group them. In computer vision, we would similarly like grouping algorithms to be based on these ecological statistics. The Bayesian framework provides a rigorous approach to exploiting this knowledge in the form of prior probability distributions. Our database enables the empirical measurement of these distributions.

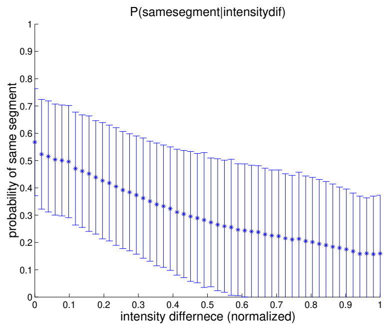


Figure 11: Similarity: The probability that two points belong to the same segment given their absolute difference in intensity (256 gray levels). We sample 1000 points from each segmentation and compute all pairwise similarities. Error bars show $\pm\sigma$ intervals.

In this section, we present our measurements of the probability distributions associated with the Gestalt cues of proximity, similarity of intensity, and convexity of regions. As another interesting empirical finding, we determine the frequency distribution of region areas and show that it follows a power law.

5.1. Proximity Cues

Experiments have long shown that proximity is an important low-level cue in deciding how stimuli will be grouped. We characterize this cue by estimating the probability that two points in an image will lie in the same region given their distance on the image plane. The results are summarized in the form of a histogram where each bin counts the proportion of point-pairs in a given distance range that lie within the same segment as designated by the human segmentor. We would like our estimate to be invariant to the granularity at which a particular image has been segmented. To this end, we scale all distances by $\sqrt{\frac{\text{number of segments}}{\text{image area}}}$.

Results are show in Figure 10. As might be expected the probability of belonging to the same group is one when the distance is zero and decreases monotonically with increasing distance.

5.2. Similarity Cues

Using a similar methodology to §5.1, we examine the probability that two points lie in the same region given their similarity. We evaluate point-wise similarity based on the absolute difference in pixel intensity (256 gray levels). This could be clearly be extended to make use of color or local texture. The results are shown in Figure 11. If images of objects were uniform in intensity over the extent of the object with some additive noise and each object in a given scene had a unique intensity, we would expect to see a curve that started at 1 and quickly decayed to 0. However, images of natural objects feature variation in intensity due to texture, shading, and lighting so the histogram we compute starts at 0.6 and monotonically decays to 0.2. This suggests that although similarity in intensity isn't a perfect cue, it does capture some useful information about group membership.

5.3. Region Convexity

One commonly posited mid-level grouping cue is the convexity of foreground object boundaries. We capture the notion of convexity for discrete, pixel-based regions by measuring the ratio of a region's area to the area of its convex hull. This yields a number between zero and one where one indicates a perfectly convex region. Since the regions in our dataset have no labels that designate them as foreground or background we are forced to look at the distribution of the convexity of all image regions. This on its own is arguably instructive and we imagine that since there can be many foreground groups and only a few background groups in a given image, the distribution for only foreground regions would look very similar. Figure 12 shows our results. As expected, grouped pixels commonly form a convex region.

5.4. Region Area

The authors of [1] approach the problem of estimating the distribution of object sizes in natural imagery by automatically finding connected components of bilevel sets and fitting the distribution of their areas. Our results from §5.2 suggest that intensity bilevel sets are only a rough approximation to perceptual segments in the image. Figure 13 shows the distribution of region areas in our data set. We get an excellent fit from a power law curve of the form $y = \frac{A}{x^{\alpha}}$ yielding an $\alpha = 1.008$.

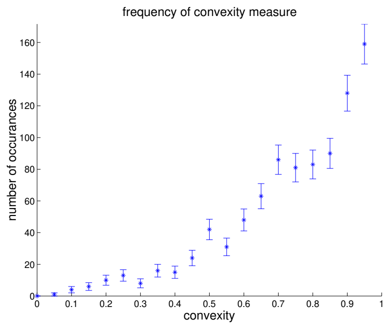


Figure 12: Convexity: The distribution of the convexity of segments. Convexity is measured as ratio of a region's area to the area of its convex hull yielding a number between 0 and 1. Error bars show $\pm\sigma$ intervals.

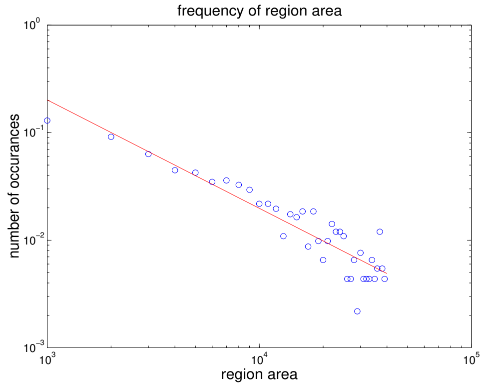


Figure 13: Region Area: This log-log graph shows the distribution in region areas. We fit a curve of the form $y = \frac{A}{x^\alpha}$ yielding an $\alpha = 1.008$. For the purposes of fitting, we throw out those sparsely populated bins which contain regions that are greater than 25% of the total image area.

6. Summary and Conclusion

In this paper, we presented a database of natural images segmented by human subjects along with two applications of the dataset. First, we developed an image segmentation benchmark by which one can objectively evaluate segmentation algorithms. Second, we measured ecological statistics related to Gestalt grouping factors. In time, we expect the database to grow to cover 1000 images, with 4 human segmentations of each image in both grayscale and color. This data is to be made available to the community in the hope that we can place the problem of image segmentation on firm, quantitative ground.

Acknowledgments

We would like to thank Dave Patterson for his valuable input, particularly in the data collection and benchmark portions of this paper. We also graciously thank the Fall 2000 students of UCB CS294, who provided our image segmentations. This work was supported in part by the UC Berkeley MICRO Fellowship (to CF), the NIH Training Grant in Vision Science T32EY 07043-22 (to DT), ARO contract DAAH04-96-1-0341, the Digital Library grant IRI-9411334, Defense Advanced Research Projects Agency of the Department of Defense contract DABT63-96-C-0056, the National Science Foundation infrastructure grant EIA-9802069, and by a grant from Intel Corporation. The information presented here does not necessarily reflect the position or the policy of the Government and no official endorsement should be inferred.

References

[1] L. Alvarez, Y. Gousseau, and J. Morel. Scales in natural images and a consequence on their bounded variation norm. In *Scale-Space Theories in Computer Vision*, 1999.

[2] J. August and S. Zucker. The curve indicator random field: Curve organization via edge correlation. In K. L. Boyer and S. Sarkar, editors, *Perceptual Organization in Artificial Vision Systems*, pages 265–287. Kluwer, 2000.

[3] S. Borra and S. Sarkar. A framework for performance characterization of intermediate-level grouping modules. *PAMI*, 19(11):1306–1312, Nov. 1997.

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

[5] R. W. Buccigrossi and E. P. Simoncelli. Image compression via joint statistical characterization in the wavelet domain. *IEEE Trans. on Image Proc.*, 8(12):1688–1701, Dec. 1999.

[6] C. Carson, M. Thomas, S. Belongie, J. M. Hellerstein, and J. Malik. Blobworld: A system for region-based image indexing and retrieval. *Third International Conference on Visual Information Systems*, Jun. 1999.

[7] O. Chapelle, P. Haffner, and V. N. Vapnik. Support vector machines for histogram-based image classification. *IEEE Trans. on Neural Networks*, 10(5):1055–1064, Sep. 1999.

[8] M. D. Heath, S. Sarkar, T. Sanocki, and K. W. Bowyer. A robust visual method for assessing the relative performance of edge-detection algorithms. *PAMI*, 19(12):1338–1359, 1997.

[9] J. Huang. *Statistics of Natural Images and Models*. PhD thesis, Brown University, May 2000.

[10] J. Huang, A. B. Lee, and D. Mumford. Statistics of range images. *CVPR*, pages 324–331, 2000.

[11] J. Huang and D. Mumford. Statistics of natural images and models. *CVPR*, pages 541–547, 1999.

[12] S. Konishi, A. L. Yuille, J. Coughlan, and S. C. Zhu. Fundamental bounds on edge detection: an information theoretic evaluation of different edge cues. *CVPR*, pages 573–579, 1999.

[13] A. B. Lee and D. Mumford. Scale-invariant random-collage model for natural images. In *Proc. IEEE Workshop on Statistical and Computational Theories of Vision*. 1999.

[14] J. Malik, S. Belongie, T. Leung, and J. Shi. Contour and image analysis for segmentation. In K. L. Boyer and S. Sarkar, editors, *Perceptual Organization for Artificial Vision Systems*, pages 139–172. Kluwer, 2000.

[15] D. L. Ruderman. The statistics of natural images. *Network*, 5(4):517–548, 1994.

[16] D. L. Ruderman. Origins of scaling in natural images. *Vision Research*, 37:3385–3395, 1997.

[17] J. Shi and J. Malik. Normalized cuts and image segmentation. *PAMI*, 22(8):888–905, Aug. 2000.

[18] J. H. van Hateren and A. V. der Schaaf. Independent component filters of natural images compared with simple cells in primary visual cortex. *Proc. R. Soc. Lond.*, 265:359–366, 1998.