

PARTS, OBJECTS AND SCENES:
COMPUTATIONAL MODELS AND PSYCHOPHYSICS

by

Laura Lynn Walker Renninger

B.S. (Massachusetts Institute of Technology) 1997

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Vision Science

in the

GRADUATE DIVISION

of the

UNIVERSITY OF CALIFORNIA, BERKELEY

Committee in charge:

Professor Jitendra Malik, Chair
Professor Martin Banks
Professor Stephen Palmer

Fall 2003

PARTS, OBJECTS AND SCENES:
COMPUTATIONAL MODELS AND PSYCHOPHYSICS

© 2003

by

Laura Lynn Walker Renninger

Abstract

PARTS, OBJECTS AND SCENES:
COMPUTATIONAL MODELS AND PSYCHOPHYSICS

by

Laura Lynn Walker Renninger

Doctor of Philosophy in Vision Science

University of California, Berkeley

Professor Jitendra Malik, Chair

In this thesis, I develop computational models to account for several different visual phenomena: (i) perception of object parts, (ii) eye movements when learning new objects, (iii) perceived shape similarity, and (iv) rapid scene identification. The validity of each model is tested using human psychophysical techniques.

Beginning with parts, I note an ecological fact: parts of objects tend to be convex. Several elaborate rules have been proposed to account for our perception of parts, but they can all be understood as an attempt to exploit convexity. I propose a model that finds convex subregions within the bounding contour of the object. The segmentations produced by the model are quantitatively compared with a large-scale data set of human object segmentations using a precision-recall framework. The model produces results

within the error range of the subject to subject variability, demonstrating that a simple convexity rule can account for human perception of parts.

It has been suggested that we use parts to encode and retrieve object memories. By studying eye movements as observers learn new objects, one can investigate the strategies that they adopt, and what information is useful for encoding object memories. I argue that, in fact, observers employ a strategy of sequential information maximization to reduce uncertainty about the orientations of the object contour. I collect eye movement data as subjects learn novel object silhouettes, and compare it to the fixations of a biologically motivated dynamic model. Model fixations are drawn away from predictable (straight) contours and toward angles near points of high curvature, similar to observed human behavior. The model collects information too efficiently, however. By adjusting the parameters until we match human performance, we can probe the limits of human sensitivity.

Objects may be recognized over successive fixations of their parts, or by matching their overall shapes, as implicitly suggested by the eye movement model. Matching objects would be straightforward if we had a perceptual shape metric that could capture the perceived similarity between two shapes. I examine two measures from the statistics and mathematics literature and ask whether they can represent human just-noticeable-differences in shape space. I find that shape discrimination thresholds are stable when measured with these metrics, for both systematic and random shape changes. This suggests that the metrics can be useful for gauging perceptual similarity of two closely related shapes.

Finally, I consider the phenomenon of rapid scene identification. Subjects are able to get the gist of a scene within a single fixation. I propose a model that learns to categorize scenes based only on the responses they evoke in V1-like filters. The model performs above chance, demonstrating that categorization could begin as early as V1. When compared with human performance on a scene identification task, the model performs much like observers who had between 37 and 50 ms exposure to the image.

Professor Jitendra Malik
Dissertation Committee Chair

In memory of
Russell L. De Valois
(1926 - 2003)

Contents

List of Figures	v
List of Tables	x
Acknowledgements	xi
I Introduction	1
II Perception of Parts	4
2.1 Introduction	4
2.1.1. Background	4
2.1.2. Summary of Approach	7
2.2 Model	8
2.2.1. Parts are Convex Subregions	8
2.2.2. Convexity Cue	8
2.2.3. Segmentation into Parts	9
2.2.4. Computational Considerations	11
2.3 Experimental Methods	11
2.4 Comparing Segmentations	17
2.5 Results	20
2.6 Future Work	23

III	Building Object Representations	25
3.1	Introduction	25
	3.1.1. Background	25
	3.1.2. Efficient Information Processing	27
	3.1.3. Summary of Approach	28
3.2	Recognizing Known Objects	29
	3.2.1. Experimental Methods	29
	3.2.2. Results and Conclusion	31
3.3	Learning New Objects	32
	3.3.1. Experimental Methods	32
	3.3.2. Computing Fixations	36
	3.3.2. Results and Discussion	36
3.4	Model	37
	3.4.1. Representing Information	37
	3.4.2. Cortical Magnification Factor	39
	3.4.3. Updating Information Over Fixations	40
	3.4.4. Groundtruth Orientation Edge Map	41
3.5	Results and Discussion	42
IV	Perceptual Shape Metrics	48
4.1	Introduction	48
	4.1.1. What is Shape?	48
	4.1.2. Quantitative Shape	49
4.2	Shape Space	49
	4.2.1. Shape of Natural Forms	49
	4.2.2. Procrustes' Distance	51
	4.2.3. Kendall's Shape Space	51
4.3	Validating the Metrics	52
	4.3.1. Experimental Methods	53
	4.3.2. Results and Discussion	57

4.4	Using the Metrics	60
	4.4.1. Experimental Methods	60
	4.4.2. Results	60
4.5	Summary and Future Work	62
V	Rapid Scene Identification	64
5.1	Introduction	64
	5.1.1. Background	64
	5.1.2. Texture as a Holistic Cue	66
	5.1.3. Summary of Approach	67
5.2	Experimental Methods	67
	5.2.1. Methods	67
	5.2.2. Constructing Confusion Matrices	71
5.3	Texture Model	73
	5.3.1. Learning Universal Textons	74
	5.3.2. Activity in Texton Channels	76
	5.3.3. Classifying New Scenes	77
5.4	Results	79
	5.4.1. Correct Superordinate-Level Categories	79
	5.4.2. Correct Basic-Level Categories	80
	5.4.3. Identification Errors	83
5.5	Discussion	86
5.6	Future Work	87
5.7	Summary	88
	References	89
A	Computing Confusion Matrices	98
B	Confusion Matrices	101

Figures

2.1	(a) Minima Rule: Points of minimum curvature are good places to begin a part cut. (b) Limbs: Part cuts are made between two points of minimum curvature when there is evidence for “good continuation”. (c) Necks: Part cuts are made between two points of minimum curvatures when a circle can be inscribed within the object that includes the two points. (d) Short-Cut Rule: All else being equal, a part cut is made from a point of minimum curvature to the nearest boundary point, crossing a local axis of symmetry.....	7
2.2	The connection between two points i and j is equal to one if the two points lie in the same convex subregion as defined by convexity, and zero otherwise.....	9
2.3	Example of Snodgrass objects and their computed contours used in the segmentation experiments.....	12
2.4	Some human segmentations. Part-lines are drawn in gray, object boundaries in black. The radius of the data circle indicates the popularity of that boundary point in segmenting the object. The colors correspond to the type of curvature of that boundary point, e.g. local minimum (red), local maximum (green), inflection point (blue). Local minima (minima rule) were used more in more than 80% of all part-lines (De Winter & Wagemans, 2001).....	13
2.5	Results from one subject using the java segmentation tool. Colors are used to highlight the regions corresponding to perceived parts.....	14
2.6	This elbow figure has multiple plausible parses. Once one parse is made, no further perceptual parses are available. This is an example of mutually exclusive parses.....	15
2.7	Some segmentations produced by the convex subregions algorithm.	

(Artifacts and “bleeding” are evident near the part lines due to our approximation using supernodes.) The segmentations are qualitatively similar to human data.....	16
2.8 Categorization of pixel pair assignments. True positive (TP), false positive (FP), false negative (FN) and true negative are also referred to as hit, false alarm, miss and correct rejection, respectively.....	18
2.9 The dots (blue) are the precision and recall measured when segmentations from each subject are compared against all other subjects as ground-truth. Because subject A is compared to subject B and subject B is compared to subject A, the dots will be symmetric about the precision = recall axis. The open circles (red) are the precision and recall measured when the segmentation based on convexity is compared to each ground-truth subject segmentation. (a) Example data for a very simple contour. Subjects show high agreement with precision and recall both near 1. The segmentation based on convexity is also good, but not as consistent as the subject-to-subject comparisons. (b) Example data for a more complex contour. Subject agreement is now much more scattered. The data from the convexity model falls within subject errors. (c) Example data from 5 subjects on 88 images. From this figure it is difficult to gauge the fit of the overall convex subregions algorithm as a of segmentation of these contours, but it is clear that the data fall within the range of subject-to-subject comparisons.....	21
2.10 The left plot shows $200 \times 200 \times 88 = 3,520,000$ F-measure values computed from segmentation comparisons between the subjects in the De Winter et. al. study. The median (red line) of this distribution is 0.84. The plot on the right shows $200 \times 1 \times 88 = 17,600$ F-measure values computed from segmentation comparisons between the same subjects and our convexity model. The median of this distribution is 0.81. See the text for a discussion.....	23
3.1 Subjects were asked to view the painting “Unexpected Visitor” under different conditions. (a) Free viewing. (b) Estimate the economic level of the people. (c) Judge their ages. (d) Guess what they had been doing before the visitor’s arrival. (e) remember the clothes worn by the people..	26
3.2 Example of fixation patterns for the four mask locations a) none b) front c) middle and d) end. Green, yellow and red dots represent early, intermediate and late eye positions, respectively.....	31
3.3 Novel objects are constructed by superimposing the inverted silhouettes of two objects from the Snodgrass and Vanderwart dataset (1980).....	33

3.4	Percent of total dwell time as a function of fixation number. Learning phases are in the top row and recognition phases are in the bottom row. The red (\wedge), green (\square), blue (\circ) and magenta (\vee) curves represent phases 1 through 4 respectively. See the text for a discussion.....	38
3.5	Schematic of the model observer for sequential information maximization. A polar grid based on the human cortical magnification factor is centered at a possible fixation f_i . The entropy in bin B is computed based on the all of the groundtruth edge orientations in the bin. The probability distribution of orientations at an edge point is updated to the bin distribution when the bin entropy is less than the current edge point entropy. The fixation that maximizes information gain is chosen as the next fixation.....	41
3.6	Example of fixation locations for the three human observers. The path based on sequential information maximization is on the bottom right.....	43
3.7	Summary of information acquisition for subject nsr. The curves for individual trials (50) during each testing phase are traced with a dotted line. The solid line represents the average behavior for that experimental condition. Error bars represent the standard deviation of the mean.....	44
3.8	The solid curves represent the average amount of information gathered during the learning phases for each subject. The dotted curves are for the recognition phases. Subjects gather information at nearly the same rate in learning and recognition trials. In the lower left panel, we can compare the curves between subjects and the model observer.....	45
4.1	Attneave's sleeping cat. The sketch is made by drawing straight lines between extrema of curvature on the image of a sleeping cat.....	50
4.2	The top shape is the reference shape. One of the test shapes (right) is identical to the reference shape and the other (left) has been altered by an affine transformation, exaggerated for illustration purposes.....	55
4.3	Set of reference shapes that were used to measure jnds for affine deformations. The shapes are ordered from left to right, top to bottom, by average $Kdist$ threshold measured from the observers. There are no obvious properties that would lead to increasing thresholds in this series..	59
4.4	Percent change in thresholds to shape perturbations as shapes to be compared are increasingly different in size.....	61
4.5	$Kdist$ thresholds increase as the shape is rotated in the plane. $Pdist$ do not appear to increase in these measurements.....	62

5.1	Pictured here are some example images from the ten scene categories used in this chapter. Each row is labeled with its basic-level (left) and superordinate-level category (right). This dataset is available at http://www.cs.berkeley.edu/projects/vision/shape	69
5.2	Subjects were shown grayscale scenes for 37, 50, 62 or 69ms followed by a jumbled scene mask and two word choices. The 2AFC task was to select the word that best described the target.....	70
5.3	Our model uses this filterbank to estimate texture features at each pixel in the image. The 36 filters consist of 2 phases (even and odd), 3 scales (spaced by half-octaves), and 6 orientations (equally spaced from 0 to π). Each filter has 3:1 elongation and is L_1 normalized for scale invariance...	75
5.4	(a) The 100 texture features found across the training images (sorted by increasing norm). These “universal textons” correspond to edges and bars of varying curvature and contrast. (b) Each pixel in an image is assigned to a texton channel based on its corresponding vector of filter responses. The total activity across texton channels for a given image is represented as a histogram. (c) Test images are categorized by matching their texton histograms against stored examples. The χ^2 similarity measure indicates our test image is more similar to a bedroom than a beach in this case.....	78
5.5	Subject accuracy in the 2AFC scene discrimination task improves with increased presentation time. The mean percent correct and standard deviation is plotted for 48 subjects (11, 15, 8 and 14 subjects at 37, 50, 62 and 69ms). Chance performance is 50% correct. At 69ms, accuracy is near 90%, confirming that the gist of a scene can be processed within one fixation.....	80
5.6	Superordinate-level confusion matrices for subjects and the model (Appendix B) are illustrated with gray levels. The order of the scene categories from top to bottom, left to right is: natural/outdoor, man-made/outdoor and man-made/indoor. Correct categorization occurs along the diagonal, which will be white for perfect performance. The amount of misclassification is represented in the off-diagonal blocks. At the superordinate-level, the model performs similar to subjects with 37-50ms of image exposure.....	81
5.7	Percent correct classifications are plotted versus basic-level scene categories, sorted by model performance. To allow direct comparison of the model with subjects, the 2AFC data has been recomputed as 10AFC data – chance performance is 10%. The dotted curves represent standard	

error measures for the model. (a) The model performs the same or better than subjects with 37ms of exposure, with the exception of mountain scenes. (b) When viewing time is increased to 50ms, the model is still able to account for subject performance on more than half of the scene categories. (c & d) The model cannot account for subject performance with more than 70ms of exposure.....

82

5.8 Plotted here is the response activity in different “scene channels” for subjects and the model. Across the x-axis of each plot are the ten scene categories, and the bars are colored according to their superordinate category (see the key). In each column, one scene category has been shown. Moving down rows, subjects view that scene for 37, 50, 62 or 69ms. A star represents the activity in the correct channel. The bottom row is response activity based on our texture analysis of the scene. See the text for a discussion.....

85

Tables

4.1	Mean thresholds measured for four naïve subjects. The thresholds for measured in <i>Kdist</i> appear more consistent for jittered and affine transformed shapes than those measured with <i>Pdist</i> . The number in parentheses are the number of quadrilaterals (and independent staircases) tested to obtain that threshold.....	57
4.2	P-values for a t-test of the null hypothesis that thresholds are the same for random jitter and systematic affine distortions of a shape for each metric....	58
4.3	95% Confidence intervals for the measured threshold values.....	58
4.4	Correlation between subjects' thresholds for different reference shapes.....	59
A.1	Sample data for a hypothetical 2AFC scene matching task. The three scene classes are A, B, and C. They would each be shown with the matching word choice and the word choice in the second column. The third column is the percentage of trials in which the subjects choose the correct word choice.....	100
A.2	Confusion matrix for three scene classes, converted from the sample 2AFC data in Table A.1.....	100
B.1	Basic-level confusion matrices estimated from 2AFC data for subjects and the texture model.....	102
B.2	Superordinate-level confusion matrices for subjects and the texture model. Collapsing across the basic-level categories yields these matrices.....	103

Acknowledgements

I would not have completed this work without the encouragement and support of many people. The most obvious, of course, is my research advisor, Jitendra Malik. Jitendra has a great sense of adventure. I was a chemical engineer who knew nothing about vision, psychology or computer science, yet Jitendra accepted me into his computer vision group to tackle vision research with psychophysics and modeling. After four years, I am officially a vision scientist, and we have managed to turn the graphics annex of the computer science department into a mini-psychophysics laboratory. Thank you for taking a chance on me.

I would also like to thank my other committee members. Marty Banks was the first to welcome me into the vision science program, and the first to sign this dissertation as I leave! The interim was filled with great advising and a warm welcome into his lab group's activities. Without his support, I surely would have felt disconnected from the researchers in Minor Hall.

Steve Palmer wrote the book on vision science (literally), and gave me my first understandable, much needed, comprehensive instruction of the psychologist's approach

to vision science. He also gave an incredible amount of time to consider this dissertation, for which I am overwhelmingly grateful.

There are several other professors that have made a lasting impact on my time here at Berkeley. Karen De Valois is legendary in her ability to lend emotional support to graduate students – especially before their qualifying exams. Cliff Schor can always be counted on for help with an eye movement study... or a bike ride. Stan Klein has been a tremendous support, from my first small research project and conference presentation, to my search for a postdoctoral position. Finally, there is Russ De Valois, to whom I've dedicated this dissertation. Russ was a pioneer in our field and a leader on campus, who still found time to challenge first year students in the classroom and host evening seminars at his home. We all miss him.

Although I sometimes felt like the “odd duck” in our group, the other kids in 545 were always there for the coffee train, sweaty burritos and whatever crazy off-track discussion struck our fancy. Alex, Alyosha, Andras, Andrea, Charless, D-Martin, Dr. Mori, Hao, and Xiaofeng: thank you for sitting through my experiments (yes, even the ones you dozed off during), tolerating the periodically missing table and giving me Linux tutorials. I hope we will see each other in Sarasota sometime!

The lovely ladies of the vision science program cheered on the completion of this dissertation, and provided all kinds of emotional support along the way. Temina and Ahna: I will miss our “journal club” meetings at Yali's each week. Laura and Kim: whether it's vision science, book club, cooking club, (what's next?), I know we will find a way to continue our friendships as we journey onto new things.

I will get into a lot of trouble if I forget to acknowledge the role my best friend played in the events leading up to this dissertation. Neil, the past ten years have been a wonderful journey, filled with surprises and challenges. You've been there beside me through it all, and I still can't believe how lucky I am. I know you were dreading the day we would no longer be Dr. and Mrs. Renninger, but I'm sure we'll find new ways to laugh together as Drs. Renninger.

And last, but certainly not least, I must acknowledge the people who knew me before I even found science. Mom, you always told me I could be anything I wanted. Without your encouragement and the demand that I finish my MIT essay, I might have missed out on the wonderful world education has revealed to me. Dad, you said "win, lose or draw, I'll always love you." As nervous as that statement made me before music competitions, I am so thankful for your unconditional love and support. It has allowed me to take risks and achieve things I might not have thought possible.