

Chapter 2

Perception of Parts

2.1 Introduction

2.1.1. Background

It was proposed long ago that we recognize shapes and objects through structural representations of “parts” and their spatial configurations (Binford, 1971; Palmer, 1975; Marr & Nishihara, 1978; Biederman, 1987). These representations are efficient for storing and recognizing objects, especially articulated objects that may appear in several different poses. In addition to studying part structure directly (e.g., Palmer, 1977), researchers have assumed some kind of part decomposition when studying category learning (Goldstone, 2000; Schyns & Rodet, 1997), figure-ground assignment and symmetry detection (Baylis & Driver, 1995) and perception of transparency (Singh & Hoffman, 1998).

How do we know parts are perceptually meaningful? Well, our perception of them is spontaneous, perhaps even preattentive. Parts often have simple names associated with them. For example, the human body is made up of arms, legs, feet,

hands, head, etc. Recent attention-based experiments have demonstrated a part superiority effect similar to the object superiority effect in scenes. It is difficult divide attention between parts within a single object, suggesting that they are perceptual units (Vecera, Behrmann & Filapek, 2001; Barenholtz & Feldman, 2003).

Assuming that parts are indeed perceptual units used in shape and object recognition, how might we parse them from a two dimensional retinal image? One idea is that we match them to shape primitives such as geons (Biederman, 1987). Another is that they might emerge through grouping or segmentation processes according to natural constraints. Hoffman and Richards (1984) proposed that the natural constraint of transversality governs our perception parts. When two concave objects interpenetrate, they intersect at a contour of concave discontinuity of their tangent planes. Carried into a two dimensional image, the idea of transversality is formulated as the *minima rule*:

Minima Rule: Divide a surface into parts at loci of negative minima of each principal curvature along its associated family of lines of curvature (Figure 2.1a).

Hoffman and Richards showed that this boundary rule predicts part perception for several reversal illusions in which the figural assignment of the contour changes.

Despite its influence at the time, the minima rule failed to describe how the two dimensional shape would actually be carved into parts. Siddiqi and Kimia (1995) proposed two rules for making part cuts for silhouette shapes.

Limbs: A limb is a part-line going through a pair of negative curvature minima with co-circular boundary tangents on (at least) one side of the part-line (Figure 2.1b).

Necks: A neck is a part-line which is also a local minimum of the diameter of an inscribed circle (Figure 2.1c).

These rules improved upon the Hoffman and Richards (1984) work by explicitly formulating how the contour minima would be linked up to form part cuts. Singh, Seyranian and Hoffman (1999) noted several examples in which limbs and necks fail to predict correct part cuts, due to the limitations of using an inscribed circle. They explore the idea that cut length and the “goodness” of a part both play a key role and propose the following:

Short-Cut Rule: A part-line is the shortest line connecting two boundary points separated by an axis of local symmetry.

Singh, et al. demonstrate that the short-cut rule can predict part cuts for several simple stimuli. They are also careful to state that, strictly speaking, this rule applies with certainty only in isolation of other factors. For example, the occurrence of a minimum of curvature along the boundary may interact with the short-cut rule to define a different part-line.

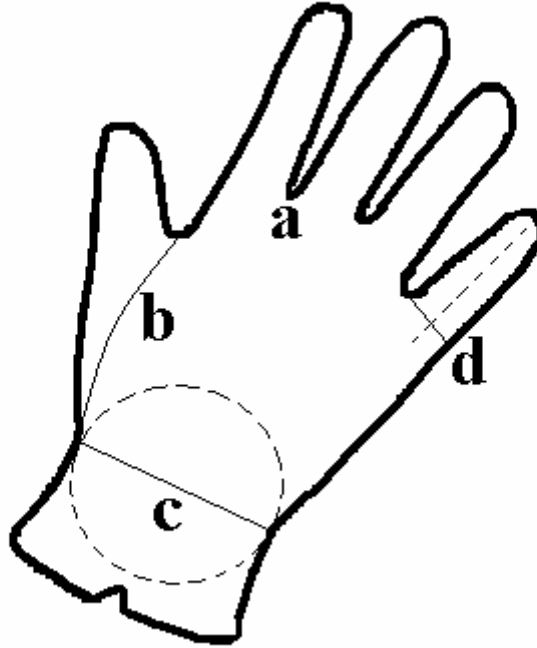


Figure 2.1. (a) Minima Rule: Points of minimum curvature are good places to begin a part cut. (b) Limbs: Part cuts are made between two points of minimum curvature when there is evidence for “good continuation”. (c) Necks: Part cuts are made between two points of minimum curvatures when a circle can 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.

It is not clear how the various rules governing part perception should be combined into a coherent framework or how they would be applied in general to arbitrary images. Also, these rules have not been quantitatively measured against a broad range of human perceptual data.

2.1.2. Summary of Approach

First, we propose a simple convexity cue and segmentation algorithm for parsing objects into parts. Second, we propose a general quantitative framework for evaluating

object segmentation algorithms and use it to measure the performance of our convexity cue. We find that convexity is a strong cue for predicting how humans segment objects into parts.

2.2 Model

2.2.1. Parts are convex subregions

It is an ecological fact that object regions in an image tend to be convex (Fowlkes, Martin & Malik, 2003). All of the rules discussed throughout the introduction attempt to capture this fundamental property indirectly. We propose a model that defines parts simply as convex subregions within an object. An object can have a hierarchy of parts, for example, a body has arms, legs and hands. Each part can also be considered as an object, but for the purpose of this experiment, we define an object to be an entity with a physical bounding contour – it is not part of any other connected object.

2.2.2. Convexity cue

Before we can begin to define parts as convex subregions, we must first define convexity. Here we adopt an intuitive and computationally simple definition. Two points lie in the same convex subregion when the straight-line path between them lies completely within the object, i.e. there is no object boundary intersecting the straight-line path between the two points. The probability that a local boundary exists (P_b), can be computed optimally from real images using brightness, color and texture (Martin, Fowlkes & Malik, 2003).

Using this definition of convexity, we now define connections between points within the object to be $1 - P_b$. In our experiments with silhouettes, the object boundaries are known, thus our connections are binary (Figure 2.2). The justification for the *minima rule* becomes obvious in our definition of convexity.

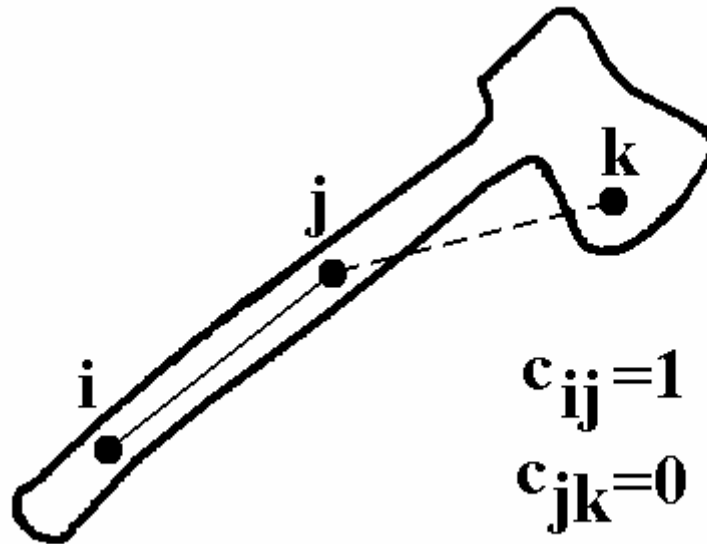


Figure 2.2. The connection c_{ij} between two points i and j is equal to one if the two points lie in the same convex subregion as defined by convexity, and zero otherwise.

2.2.3. Segmentation into Parts

Once connections are defined for every pair of pixels that lie on the object, we can segment the object into maximally convex subregions A and B by maximizing the number of connections within each region and minimizing the number of connections between regions. This can be accomplished by minimizing the normalized cut criterion. (Shi & Malik, 2000).

$$Ncut(A, B) = \frac{cut(A, B)}{assoc(A, A \cup B)} + \frac{cut(A, B)}{assoc(B, A \cup B)}$$

$$cut(A, B) = \sum_{i \in A, j \in B} c(i, j)$$

$$assoc(A, A \cup B) = \sum_{i \in A, k \in A \cup B} c(i, k)$$

The term $c(i, j)$ is the “connection” between two points i and j defined above. An approximate solution to minimizing the normalized cut criterion can be found by solving the eigensystem $(D - C)x = \lambda Dx$, where C is the symmetric $N \times N$ matrix of connections between the N object points and D is a diagonal matrix in which each entry is the total number of connections to an individual point

$$d(i) = \sum_j c(i, j).$$

The eigenvector of the second smallest eigenvalue is a real-valued approximation of the best segmentation of the object points into two parts. Because the approximation is real-valued, some threshold or grouping of the vector values must be done to acquire the segmentation. Our convex subregions algorithm uses the k-means approach to find the two best clusters for the bipartion (Hastie, Tibshirani & Friedman, 2001).

Additional parts can be found by looking at the next largest eigenvectors, but the estimations in these vectors become less reliable. In order to maintain the best

estimations, our algorithm recursively bipartitions each region. The recursion can be stopped when some *Ncut* threshold value is reached, or when a certain number of regions have been generated. This “zooming-in” continually changes the scale at which part-lines are found, which has an intuitive correspondence with the “coarse-to-fine” theme common in visual processing.

2.2.4. Computational Considerations

The number of pixels contained within the bounding contour of an imaged object can be arbitrarily large, making the task of approximating the maximally convex subregions computationally challenging. We reduce the size of the connection matrix C by grouping pixels together into “supernodes” based on their similar connections to other pixels on the object. The connection to each resulting supernode is just the sum of total connections to each individual pixel contained in the node.

2.3 Experimental Methods

Previous work in the area of part perception has only tested theories on a limited set of stimuli, and the performance of these theories in explaining part perception is often judged subjectively by the researcher instead of objectively through a quantitative comparison with human data. De Winter and Wagemans (2001) recognized this problem and conducted a large-scale study of human part perception. They constructed a dataset of 260 object contours from a well-known set of line-drawn objects (Snodgrass & Vanderwart, 1980). The contours were corrected to prevent merging of tangential sections, etc. (Figure 2.3).

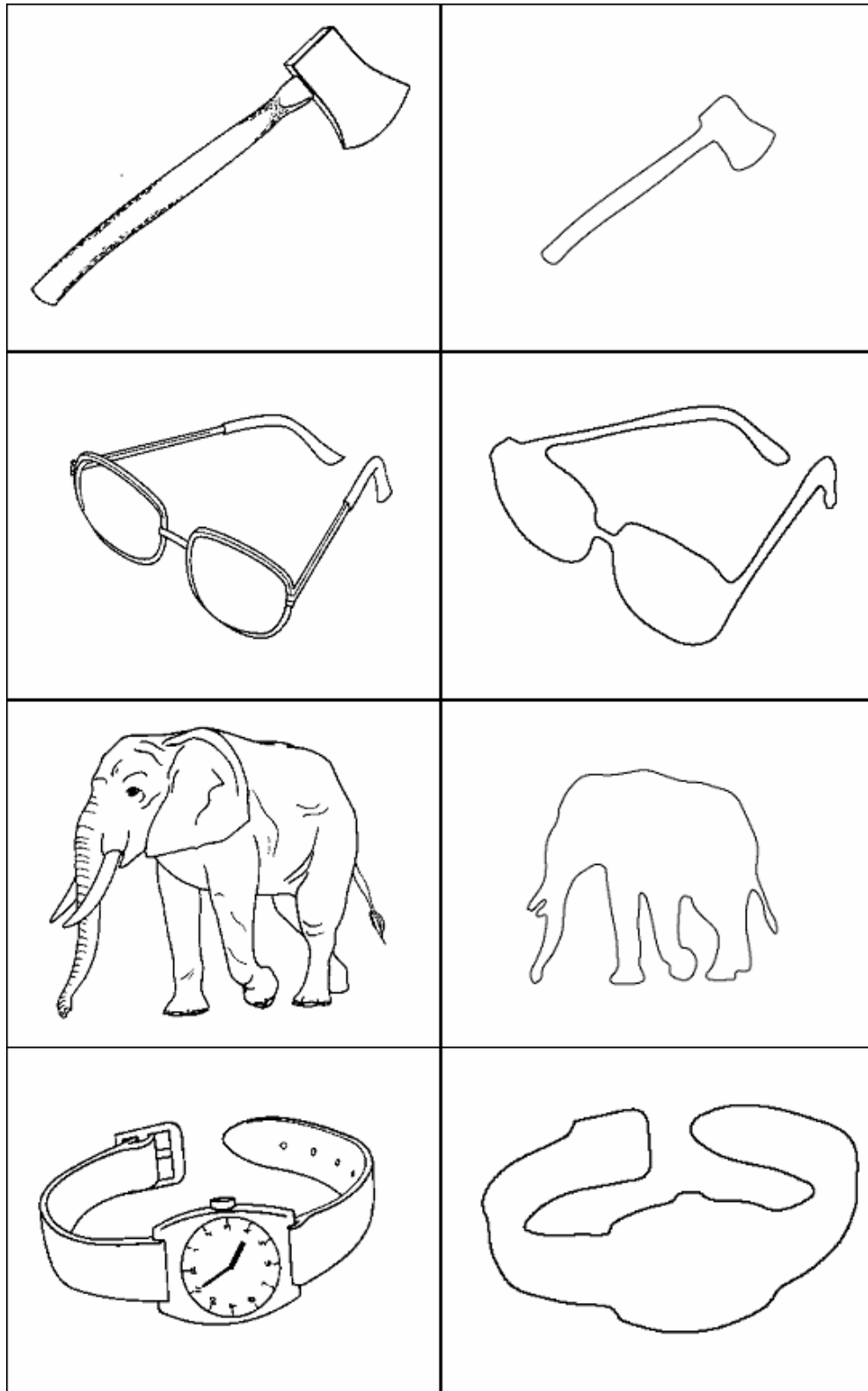


Figure 2.3. Example of Snodgrass objects and their computed contours used in the segmentation experiments.

About 200 human subjects segmented 88 of these contours into parts during a pencil and paper task, and the straight-line part cuts were recorded. De Winter et al. found examples of the many of the rules for part perception evident in their data, and they also found some examples that could only be explained by top-down influences, for example, segmentation of an arm at the elbow seems to be due to our knowledge of how joints bend (Figure 2.4).

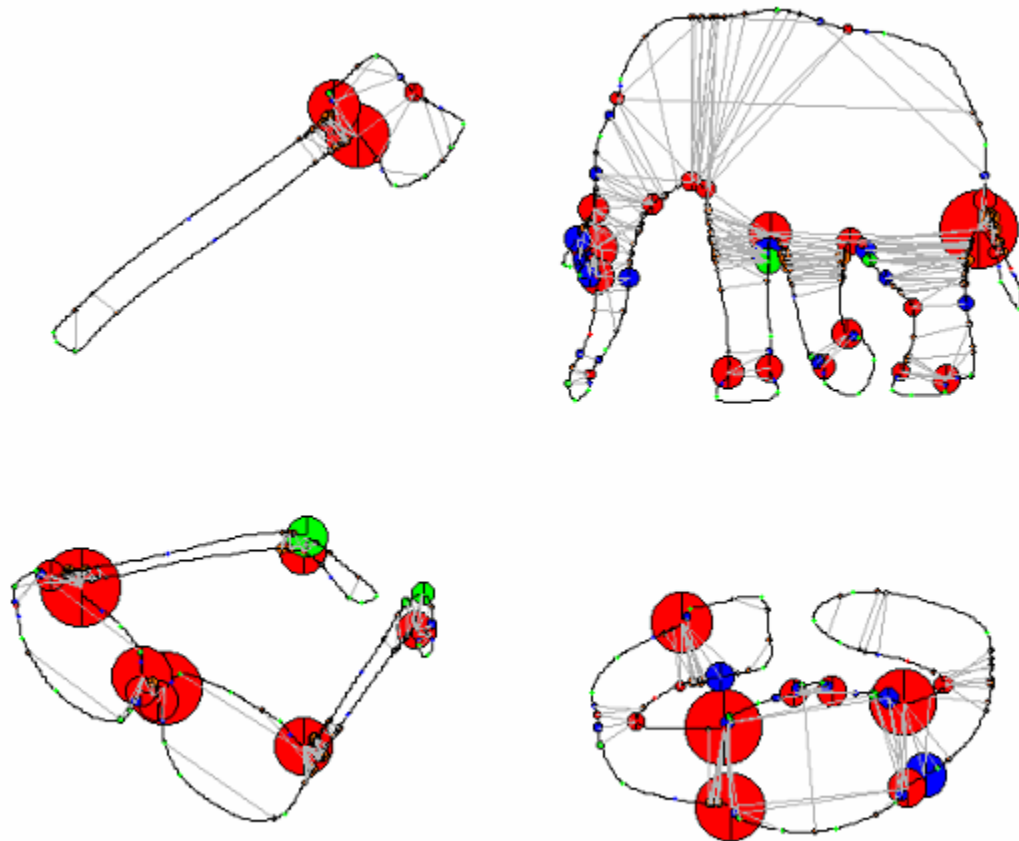


Figure 2.4. Some human segmentations. Part-lines are drawn in gray, object boundaries in black. The radius of the data circle indicates the popularity of that boundary point in segmenting the object. The colors correspond to the type of curvature of that boundary point, e.g. local minimum (red), local maximum (green), inflection point (blue). Local minima (minima rule) were used more in more than 80% of all part-lines (De Winter & Wagemans, 2001).

We expanded the human data by collecting segmentations from 9 subjects for all 260 contours (CPHS protocol 2002-5-75). Our subjects used a java segmentation tool (Martin, Fowlkes, Tal & Malik, 2001) and were allowed to make curved as well as straight-line part cuts (Figure 2.5).

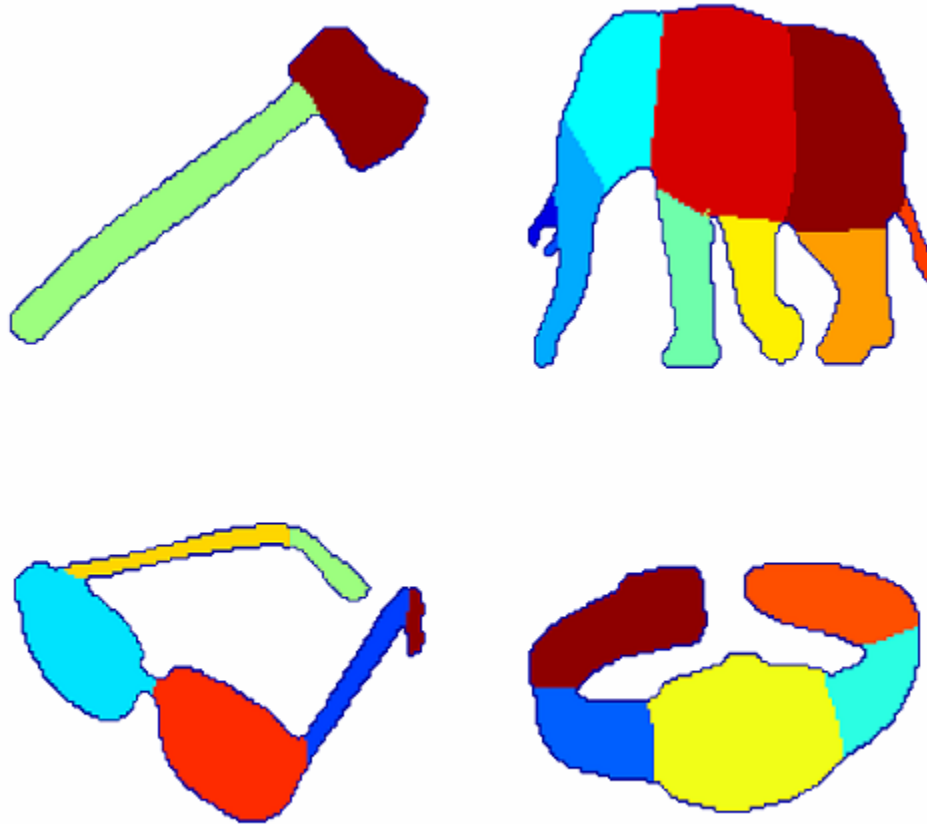


Figure 2.5. Results from one subject using the java segmentation tool. Colors are used to highlight the regions corresponding to perceived parts.

From the human data, it is clear that multiple parses are available. For example, on the watch, some subjects chose to parse the watchband where it bends and others did not. Some subjects chose to segment the body of the elephant and the feet of the

elephant. There is often a hierarchy of parts available, and differences between subjects' segmentations are due to the level of granularity (or scale) to which they segment. In other cases, there may be mutually exclusive segmentations, in which choosing a particular part-line renders others unavailable (Figure 2.6).

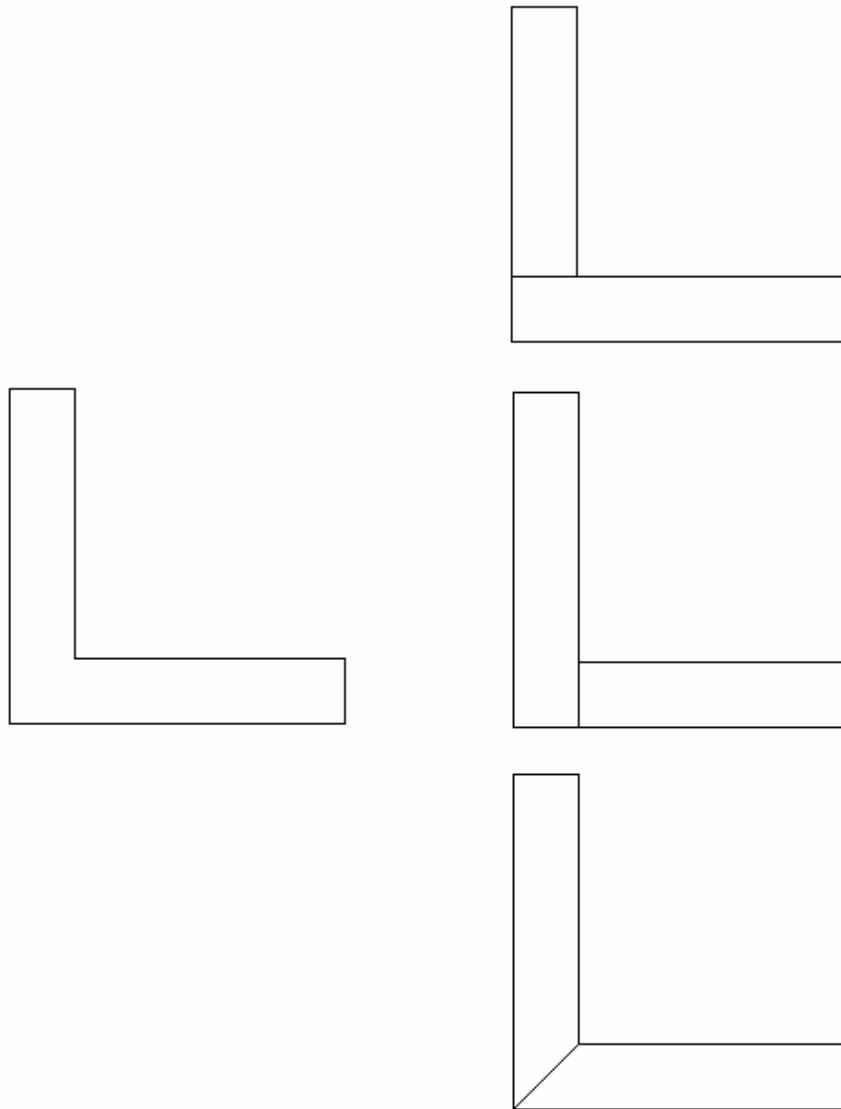


Figure 2.6. This elbow figure has multiple plausible parses. Once one parse is made, no further perceptual parses are available. This is an example of mutually exclusive parses.

Our convex subregions algorithm produces segmentations according to two parameters: how many parts to produce and a threshold value for the cut (Figure 2.7). In our experiments, we recursively cut the object until the algorithm reached a threshold of $N_{cut} = 1$ or until it produced 10 parts. When $N_{cut} = 1$, there are no “weak points” in the graph suitable for cutting because the graph is fully interconnected.

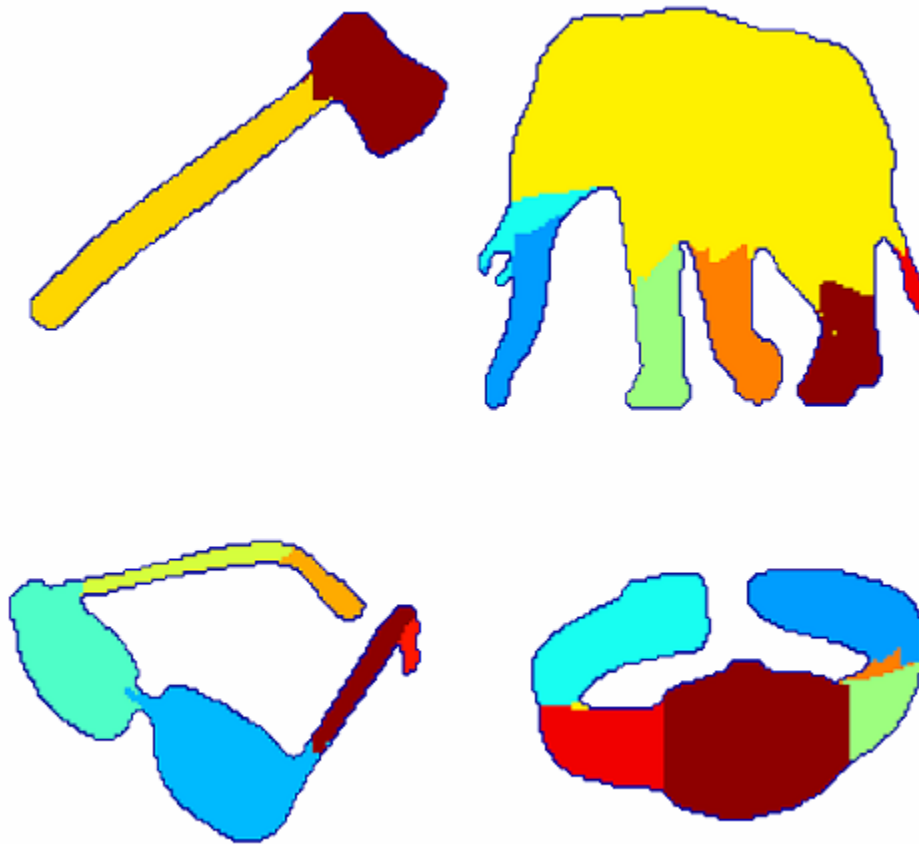


Figure 2.7. Some segmentations produced by the convex subregions algorithm. (Artifacts and “bleeding” are evident near the part lines due to our approximation using supernodes.) The segmentations are qualitatively similar to human data.

It should be noted that although we are using simplified object silhouettes in these experiments, the algorithm can be applied to any arbitrary image in which boundaries are defined, such as the output from an edge detector. The defined boundaries can be hard (binary) or soft (probabilistic) and may include internal markings and features within the object in addition to the self-occluding contour of the object.

2.4 Comparing Segmentations

Despite their large-scale study, it was difficult for De Winter et al. (2001) to quantitatively evaluate the performance of the proposed rules at predicting human part perception because (1) the outcome produced from a combination of the various rules is not always obvious or directly computable from the image and (2) they lacked a framework for quantitatively comparing segmentations. Our algorithm avoids the first issue by producing a segmentation directly from an input image using convexity as its only cue. We address the second issue by treating segmentation as a pixel classification problem.

Given a ground-truth human segmentation of the object, we can score whether or not a pair of pixels has been assigned to the same part ($S_{ij} = 1$) or to different parts ($S_{ij} = 0$). A segmentation, \hat{S}_{ij} , produced by the pixels-to-part classifier (either a different subject or our algorithm), can be evaluated against ground-truth within the precision-recall framework.

$$precision = P\left(S_{ij} = 1 \mid \hat{S}_{ij} = 1\right)$$

$$recall = P\left(\hat{S}_{ij} = 1 \mid S_{ij} = 1\right)$$

In words, precision measures how often the classifier is correct when it assigns two points to be on the same part. Of all the correct assignments, recall is the proportion of them found by the classifier.

Precision-recall curves are isomorphic to the Receiver Operating Characteristic (ROC) curves used in signal detection theory. When comparing segmentations, each pixels-to-part classification counts as either a true positive, false positive, false negative or true negative (Figure 2.8).

		S_{ij}	
		(groundtruth)	
		1	0
\hat{S}_{ij}	1	TP	FP
	0	FN	TN
(test)			

Figure 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.

In ROC analysis, *true positive detection rate* and *false positive detection rate* are measured for various classification criterion thresholds and a curve is plotted. The area under the curve provides a statistical measure of the accuracy of the classifier.

$$\text{true_positive_rate} = \frac{TP}{TP + FN}$$

$$\text{false_positive_rate} = \frac{FP}{FP + TN}$$

In the precision-recall framework, recall is precisely the *true positive rate*. Precision is also a false positive rate, but the rate is computed in terms of the total number of positive responses by the classifier (i.e. $\hat{S}_{ij} = 1$) instead of the total number of negatives in the ground-truth segmentation (i.e. $S_{ij} = 1$).

$$1 - \text{precision} = \frac{FP}{FP + TP}$$

The *false positive rate* in ROC analysis will vary from 0 to 1 while the quantity *1-precision* varies from 0 to an upper bound that depends on the number and size of parts in the ground-truth segmentation. It is possible to debate which analysis is more appropriate given a classification or a detection task but both frameworks capture a trade-off between sensitivity and selectivity and can be used to compare the relative performance of

different classifiers. Similar to using the area under a ROC curve to get a single measure of classifier accuracy, precision and recall can be combined into their weighted harmonic mean called the F-measure.

$$F = \frac{precision * recall}{(\alpha * precision + (1 - \alpha) * recall)}$$

By changing the value of α , we can specify the importance of sensitivity and selectivity in evaluating performance. In our evaluation, we set $\alpha = 0.5$. F will equal 1 when the segmentations being compared are identical, and it will decrease as the discrepancy between segmentations increases.

2.5 Results

We cannot control the threshold chosen by subjects in parsing the objects, and the nature of our algorithm (to produce bipartitions at each step) also prevents us from continuously manipulating a threshold. For this reason, each segmentation comparison produces a single point on the precision-recall plot. Figure 2.9a shows a precision-recall plot for one of the simplest contours. The subjects' segmentations agree and cluster near 1 for both measures. The segmentation predicted by the convex subregions algorithm gives a good result, but does not agree exactly with the subjects. Figure 2.9b shows a precision-recall plot for a more complex contour. Now subjects vary widely in their segmentations and the model based on convexity falls somewhere within the subject-to-subject variability. Figure 2.9c shows subject and model data for 88 of the contours. It is

difficult to gauge overall performance of the convexity model from this raw plot. For this reason, we turn to the F-measure.

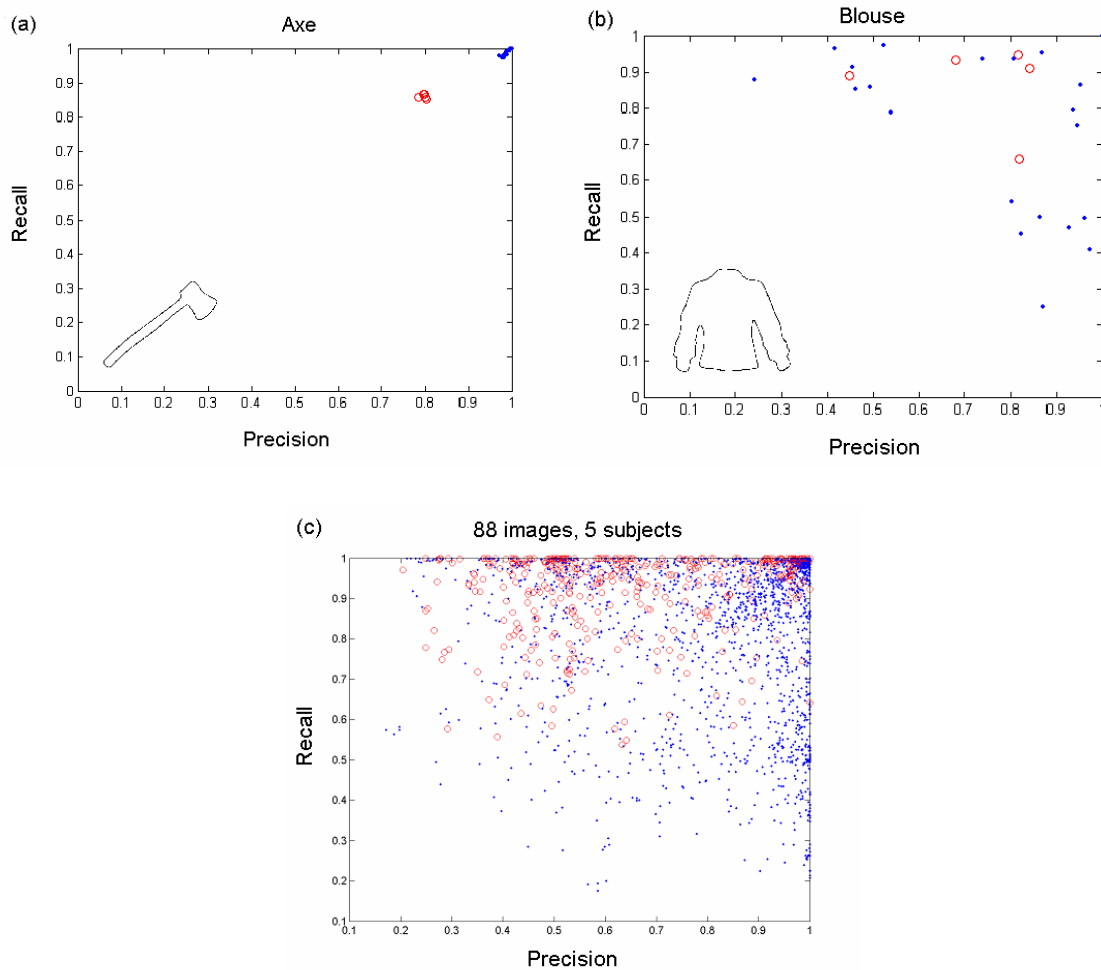


Figure 2.9. The dots (blue) are the precision and recall measured when segmentations from each subject are compared against all other subjects as ground-truth. Because subject A is compared to subject B and subject B is compared to subject A, the dots will be symmetric about the precision = recall axis. 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.

The F-measure provides a more concise look at the performance of the convexity model compared to subject segmentations. Figure 2.10 shows the distribution of all F-measures obtained when subjects from the De Winter et al. study were compared to each other and to the convexity model. The richness of the data in the left plot produces a smooth distribution, with a median of 0.84. The main peak occurs at an F-measure of one, indicating that for the most part, humans select the same segmentations. The gradual fall-off is consistent with slight deviations in placement of the part-line and in the granularity of the segmentation. A secondary peak occurs below an F-measure of 0.6, and is likely caused by alternate, or mutually exclusive segmentations as described in Figure 2.6. The right plot shows a similar distribution of F-measure values when the output of our convexity model is compared against the human data. The data in this plot represents fewer comparisons, and is therefore less smooth, however the shape of the distribution is similar with a main peak near 1 and a secondary peak just below 0.6. The median of this distribution is 0.81, which falls just below the median for human versus human comparisons, and its main peak is at 0.97, which is also just below the value of 1.0 obtained for human versus human comparisons.

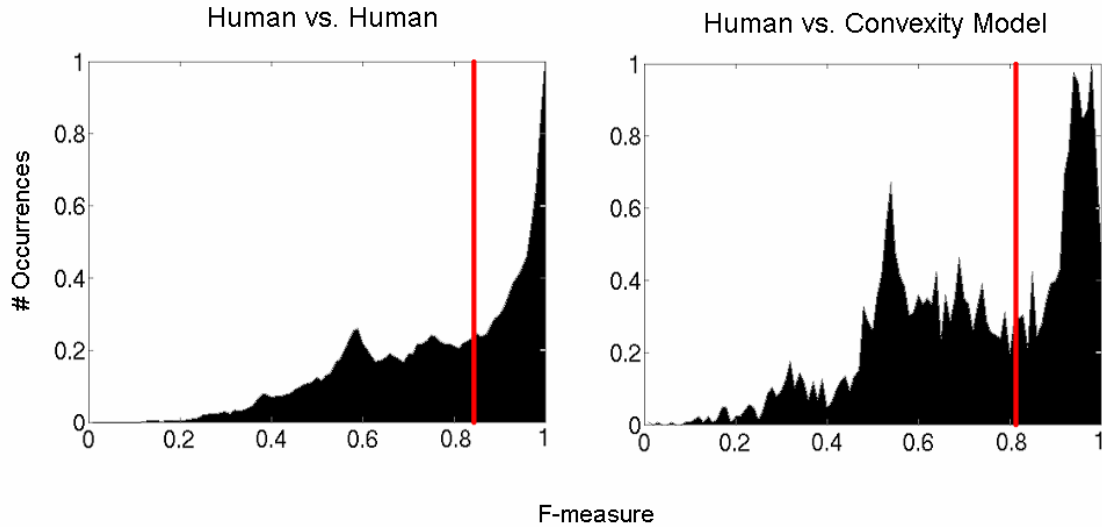


Figure 2.10. The left plot shows $200 \times 200 \times 88 = 3,520,000$ F-measure values computed from segmentation comparisons between the subjects in the De Winter et. al. study. The median (red line) of this distribution is 0.84. The plot on the right shows $200 \times 1 \times 88 = 17,600$ F-measure 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.

2.6 Future Work

It would be interesting in future work to investigate the perceptual validity of the *Ncut* criterion. To what extent does the *Ncut* value indicate the goodness of a part cut? Also, our algorithm stopped parsing the contour once there were 10 parts, or once we hit an *Ncut* value of 1. There may be a better stopping criterion.

With regards to the convexity model and segmentation algorithm, several other configurations could be tried to improve the segmentation outputs. For example, we could employ a “course-to-fine” strategy in placing part-lines. First, our supernodes would be used to get a rough region in which the part-line would occur. Next, the

original pixels would be used to place the part-line more exactly, thereby avoiding the artifacts we currently get in the output segmentation.

A second area for exploration is the cut criteria. We use the normalized cut, which was developed to specifically to penalize cutting very small regions. Often, our parts are actually quite small with respect to the entire object, so a different cut method might be more appropriate (see Katz and Tal (2003) for a different cut method used in 3D mesh segmentation).

While the use of convexity in our algorithm has done a good job in explaining much of the performance of humans, it still falls a bit short, as would be expected if subjects are using more cues than just convexity. It is already known that subjects make use of top down knowledge of parts (e.g., joints) when parsing these silhouettes. The 88 contours in the DeWinter et al. dataset are balanced with 44 being easy to recognize and 44 being difficult to recognize. By separating the performance on these subsets of data, we might get a measure of how much top-down knowledge is being used in segmenting the contours in the first subset. Unfortunately, many of the difficult to recognize objects have round-ish contours without distinctive parts. A better dataset might include novel objects versus known objects, rather than difficult to recognize objects versus known objects.

We conclude from this work that convexity is a very strong cue for parsing object silhouettes into parts and that segmentation schemes can be effectively evaluated using a benchmark dataset and the precision-recall framework.