

The Perception of Rectangular Area and Guidelines for Creating Effective Treemaps

Abstract—Treemaps are space-filling visualizations that make efficient use of limited display space. They recursively subdivide rectangular areas to encode numerical values and because of their efficiency they are commonly used to display large amounts of hierarchical data. However, area judgments are known to be less accurate than judgments of other visual encodings, such as length. We investigate the effects of rectangle aspect ratio and data density (data marks per unit area) on value comparison judgments involving both leaf and non-leaf nodes. Our study of aspect ratios finds that 90-degree rotation has little impact on estimation accuracy and, contrary to common assumptions, a wider distribution of aspect ratios results in higher accuracy. We then compare treemaps with hierarchical bar chart displays and identify the data density at which length-encoded bar charts become less effective than area-encoded treemaps. Based on these results, we propose guidelines for the effective use of treemaps and suggest alternate approaches to treemap layout.

Index Terms—Graphical Perception, Information Visualization, Treemaps, Visual Encoding, Experiment, Mechanical Turk.

1 INTRODUCTION

- Amount of data available keeps growing, and large datasets are becoming easily accessible [?, ?, ?]. Visualization is an effective tool to help identify trends or patterns, or discover a story, in the data.
- Visualizations are often touted for their space-filling properties [8, 6, 11]. Treemaps are one of the most commonly used space-filling visualization types [14, ?, ?, ?]. However, treemap design is not rooted in empirical evidence.
 - Treemaps use a non-optimal encoding [?]. Length or position is known to be a superior encoding [?, 3, ?], but is less space effective (only uses one dimension of the space). What is the tradeoff between space-savings and readability? Given a limited amount of space and the amount of data to visualize, when is it best to use what type of encoding?
 - Squarified treemaps claim to ease area comparison by minimizing aspect ratios, but this claim has heretofore not been evaluated.
- Need guidelines for choosing the most effect visualization type. Would also like recommendations that guide future design of layout algorithms etc.

2 RELATED WORK

Graphical perception. Bertin [?]: “Resemblance, order, and proportion are the three signifieds in graphics.” Previous work on comparing visual encoding variables. Talk about area in particular (Steven’s Law). Talk about Cleveland & McGill, Heer & Bostock and others. Mention integral / separable dimensions?

Heer & Bostock find that circular area, rectangular area (comparing 2 rectangles), and rectangular area (in treemaps) have similar judgment accuracy. They also find an affect due to aspect ratio in both rectangular comparison conditions. Comparing aspect ratios of 2/3, 1, 3/2, they found that comparing squares resulted in significantly higher

error. We seek to replicate this result, and extend the analysis to a wider variety of aspect ratios.

Treemap work on communicating structure. What are all the previous studies on treemaps? Kosara/Metaphor [16]: mention effects of task prompt / metaphor (“under” vs. “contained in”)? Note that we are not investigating structure directly in this work, however, we are incorporating comparisons between leaf and non-leaf nodes.

3 RESEARCH GOALS

What are we trying to achieve?

What aspects are common to each experiment? Note shared error measure.

For example, mention Mechanical Turk issues here? (Introduce MTurk as a shorthand for Mechanical Turk). Introduce notions of HIT, reward, and qualification task.

4 PILOT STUDIES

Before investigating effects due to aspect ratio or data density, we first conducted pilot studies to inform our experimental approach. Specifically, we wanted to test the effects of true proportional difference, luminance contrast, and judgment type. To do so, we conducted two separate experiments.

4.1 Pilot 1: Proportional Difference and Luminance

In our first pilot study, we showed subjects a 600×400 pixel squarified treemap display visualizing 24 uniform random values. Two rectangles were selected at random and marked **A** and **B** for comparison. We asked subjects to identify which rectangle was smaller and what percentage the smaller was of the larger. In each trial, the RGB intensity of each cell was varied randomly between 0.3 and 1.0 according to a uniform distribution.

We conducted the pilot on MTurk as 100 HIT distinct hits, each with 24 assignments and a reward of \$0.03. A total of 41 subjects provided 2400 responses; we removed 121 outlier responses (5%) with an absolute error greater than 35%. We then analyzed responses by applying Analysis of Covariance (ANCOVA) to the log absolute error of subjects’ proportion estimates, treating both true proportional difference and luminance difference as covariates.

Multiple prior studies [3, 5] have demonstrated an effect of the true proportional difference on the accuracy of proportional judgments. As shown in Figure 1, our pilot study results exhibit a nearly identical profile for rectangular area judgment as that of Heer & Bostock [5]. Our analysis finds that the true difference has a strong, statistically significant effect on accuracy ($F(1,2252) = 253.90, p < 0.001$). We also note a high prevalence of responses that were a factor of 5 and that as a result, trials for which the true difference is also a factor of

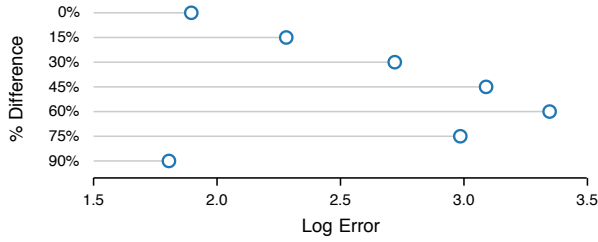


Fig. 1. Average area judgment errors in Pilot 1, binned by true difference. Errors peak at 60% and diminish at the extremes of the scale.

5 correspondingly exhibited less error. This reduced error is presumably an artifact of estimation bias and not due to improved perception. These results verify that the true proportional difference has a strong effect on judgment accuracy and should be carefully controlled across experimental conditions.

We also sought to verify that luminance differences among treemap cells—commonly used to encode an additional quantitative variable—do not interfere with area judgments. In other words, we wanted to confirm that luminance and area are *separable* perceptual dimensions [4, 13]. Our analysis finds no significant effect due to luminance difference ($F(1,2252) = 0.086$, $p < 0.767$), in agreement with prior work. The implication is that studies of area judgment do not need to include interactions with luminance in order to produce correct, generalizable results.

4.2 Pilot 2: Proportional Judgment Type

Proportional judgments of the form “what percentage is the smaller of the larger” have been used in numerous graphical perception experiments (e.g., [3, 9, 15, 5]), often to aid comparison with prior studies. However, one can equivalently express proportions in terms of multiples: as a scalar factor that is the reciprocal of the percentage. If equivalent, the ranking of visual variables (e.g., length vs. area) and systematic estimation biases (e.g., underestimation of area as predicted by Steven’s Power Law [12]) should be invariant to how the judgment task is formulated. Thus in our second pilot study, we sought to check that these two response types (percentage and scalar difference) are equivalent in terms of both bias and absolute error.

Subjects were shown either two circles (requiring an area comparison) or two rectangles of equal height (requiring a length judgment) in a 600x400 pixel image. In both cases, the shapes were center aligned. One shape was marked with the number “100” and the other with a question mark (“?”). We asked subjects to estimate the unknown value using the other shape as a reference. When the smaller shape is unknown, the question takes the form of a percentage judgment with valid responses between 0-99. When the larger shape is unknown, the question takes the form of a scale judgment indicating by what factor the unknown shape is larger. For each combination of shape and task, we tested proportional scale factors of 1.4, 1.8, 2.4, 3.6, 5.1, 8.3, 14.4, and 24.2. The study was deployed on MTurk as 32 HITS (2 judgment types \times 2 encodings \times 8 differences), each with 24 assignments and a reward of \$0.04. A total of 29 subjects provided 768 responses; we removed 7 outlier responses (0.9%) with an absolute percentage error greater than 35%.

We then analyzed responses using Analysis of Variance (ANOVA) of the log absolute percentage errors. Consistent with prior results [3, 5], we found a significant advantage for length judgments over area judgments ($F(1,729) = 26.84$, $p < 0.001$) and again found a strong effect of true proportional difference ($F(7,729) = 42.38$, $p < 0.001$). The raw (non absolute) error scores revealed that area judgments on average resulted in underestimates, again consistent with prior work [12, 10]. Finally, we found no significant effect due to the judgment type ($F(1,729) = 0.001$, $p = 0.969$). We also found similar results when analyzing error in terms of scalar, rather than percentage, differences. These results provide evidence that there is no significant performance difference between percentage and scale judgments and

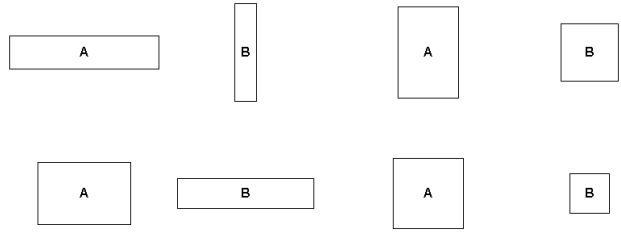


Fig. 2. Example stimuli from the aspect ratio study. Rectangles varied in terms of both proportional difference and aspect ratio.

justifies the continued use of percentage judgments tasks in subsequent experiments.

5 EXPERIMENT 1: THE EFFECTS OF ASPECT RATIO

Aspect ratio experiment and results goes here. List motivation and hypotheses here. Motivation: - high aspect ratios are bad [7, 13]. - squares? heer and bostock

Hypotheses: - difference will again be a factor - rotation will hamper comparison. Rotation is more cognitively difficult than judgments of translation and scale [1]. - large ratios? - squares?

5.1 Method

We asked subjects to compare rectangles of varying size and aspect ratio. We showed subjects a 600x400 pixel image containing two center-aligned rectangles and instructed them to identify which of the two rectangles (marked A or B) was the smaller and then estimate the percentage the smaller was of the larger by making a “quick visual judgment.” Note that stimulus images consisted solely of two rectangles (Fig. 2) and not a full treemap display. As Heer & Bostock [5] found no significant accuracy differences between these two stimulus types, we are confident that our results generalize to treemap displays.

We controlled both the true proportional difference between rectangles and their aspect ratios. True differences varied over 0.32, 0.48, 0.58, 0.72. To reduce accuracy differences due to estimation bias, these values were explicitly placed at equal, symmetric distances from their nearest factor of 5. Rectangle aspect ratios were determined by the cross-product of the set $\{\frac{2}{9}, \frac{2}{3}, 1, \frac{3}{2}, \frac{9}{2}\}$ with itself. This set was chosen to facilitate comparison with the results of Heer & Bostock [5] (who used the set $\{\frac{2}{3}, 1, \frac{3}{2}\}$) and also to probe the effects of including more extreme aspect ratios. As all non-square aspect ratios have a matching rotated variant (e.g., a rectangle with ratio $\frac{2}{3}$ is a 90° rotation of a rectangle with ratio $\frac{3}{2}$), we included an additional replication of the 1×1 condition for balance. Our experiment design thus consisted of 104 unique trials (HITS): 4 (difference) \times 26 (aspect ratios with replication).

As a qualification task we used multiple-choice versions of two example trial stimuli. For each trial, we recorded subjects’ discrimination (which rectangle was smaller) and proportion (what percentage the smaller is of the larger) judgments. We did not analyze timing data due to known issues with the standard MTurk interface [5]. We requested a total of 104 HITS with N=25 assignments and paid a reward of \$0.03 per HIT.

5.2 Results

We collected $104 \times 25 = 2,600$ responses, from which we removed 18 outliers (0.7%) with absolute errors above 35%. To analyze the data, we used log absolute error: $\log_2(|\text{judged percent} - \text{true percent}| + 1)$. We then conducted an ANOVA with a $4 \times 6 \times 2$ factorial design:

- (4) *proportional difference*: 0.32, 0.48, 0.58, 0.72
- (6) *aspect ratio pairs*: rotated variants ($\frac{2}{9}, \frac{9}{2}$) are grouped
- (2) *relative orientation*: same ($\frac{9}{2} \times \frac{9}{2}$) or different ($\frac{9}{2} \times \frac{2}{9}$).

The model is partially unbalanced, as orientation does not apply to comparisons with squares.

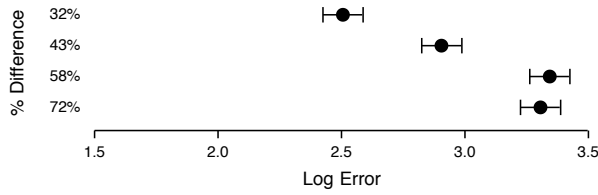


Fig. 3. Area judgment error by true proportional difference. Error bars indicate 95% confidence intervals.

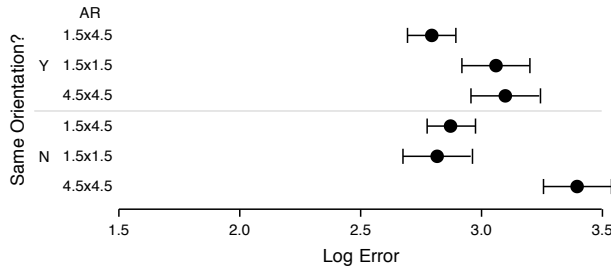


Fig. 4. Judgment error by orientation and aspect ratio. Squares omitted due to rotation invariance. Error bars indicate 95% confidence intervals.

5.2.1 Proportional Difference Dominates

We again found a strong effect due to the true proportional difference ($F(3,2173) = 94.56, p < 0.001$). True difference produced the strongest effect in our model, shifting average absolute errors by up to 18%. This result argues for the importance of including true difference as either a factor or covariate in proportional judgment studies. Applying Bonferroni-corrected post-hoc tests, we found that all difference levels were significantly different ($p < 0.05$) except for 0.58 and 0.72. Finally, the effects of proportional difference appear to be independent of the other factors; we found no significant interactions ($p > 0.05$) with either orientation or aspect ratio.

5.2.2 Orientation Affects Extreme Aspect Ratios

We then examined the effects of shared orientation on judgment accuracy, excluding comparisons involving squares. We found no main effect due to orientation ($F(1,1490) = 0.669, p = 0.414$). This result implies that, on average, 90° rotation of rectangles has little to no effect. However, we did find a significant interaction effect between orientation and aspect ratio ($F(2,1490) = 7.23, p < 0.001$). Figure 4 shows error rate by both orientation and aspect ratio. Mental rotation appears to increase error when comparing the most extreme ratios in our study (4.5 x 4.5) and suggests that rotation may contribute to higher judgment errors as aspect ratios deviate further from squares (e.g., as occurs in slice-and-dice treemaps [8]).

5.2.3 Diverse Aspect Ratios Improve Accuracy

Finally, we analyzed the impact of aspect ratio on judgment accuracy, finding a significant effect ($F(5,2173) = 13.85, p < 0.001$). Applying post-hoc tests with Bonferroni correction, we found that aspect ratio pairs of 450x450 and 100x100 exhibited significantly higher error than the pairs 100x150, 150x150 and 150x450. Similarly 100x450 was significantly more error prone than 150x450. All other differences were not significant. Figure 5 shows the resulting rank ordering by error of aspect ratio pairs and their corresponding confidence intervals. The results indicate that average judgment accuracy improves when comparing rectangles with diverse aspect ratios, even when one of the aspect ratios is large. The highest error occurred when comparing two extreme aspect ratios or comparing squares. The latter result replicates the results of Heer & Bostock [5], who similarly found increased error for comparison of squares.

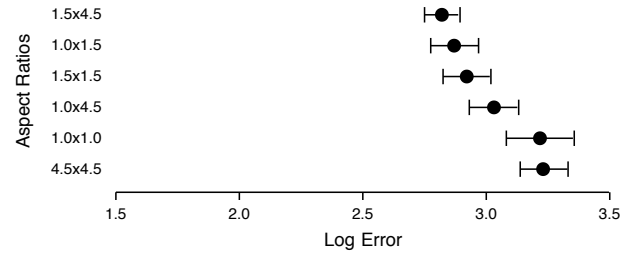


Fig. 5. Area judgment error by aspect ratio. Squares and extreme ratios have the highest error. Error bars indicate 95% confidence intervals.

5.3 Discussion

Our experiment found that graphical perception suffers when comparing large aspect ratios, particularly when the rectangles having different orientations. These results support the general intuition against treemap layout algorithms that produce rectangles with large aspect ratios (e.g., slice-and-dice [8]). On the other hand, subjects exhibited equally poor accuracy when comparing squares. As a result, the perceptual justification for squarified treemap layout algorithms [2, 14]—that squares promote more accurate comparisons—appears to be faulty. Rather, our findings are consistent with the hypothesis that subjects partially rely on 1D length comparisons when estimating area, as comparing the lengths of sides as a proxy for area leads to maximal error when comparing squares. It appears that squarified algorithms are effective in part because (a) they avoid extreme aspect ratios and (b) they are unable to perfectly achieve their “squarification” objective, instead producing a distribution of aspect ratios. As we discuss later in the paper, this insight can also be applied to inform improved approaches to treemap layout.

6 EXPERIMENT 2: THE EFFECTS OF DATA DENSITY

Our first experiment investigated value comparison accuracy in two types of charts designed to display hierarchical data: a treemap and a grouped bar chart. We used a squarified treemap layout [2], as it is the most commonly used treemap layout. In addition, our results from our aspect ratio experiment suggest that squarified treemaps produce marks that users can effectively compare. We did not ask questions about the structure of the tree, instead choosing to focus on the value comparison task, which we believe is the more common task (**NK: Probably should do some studies asking about the structure of the tree?**).

We designed this experiment to answer the following questions:

- How does choosing a treemap or grouped bar chart display affect estimation time and accuracy?
- How does data density affect estimation time and accuracy?

We were interested in three types of comparisons: leaf to leaf, leaf to non-leaf, and non-leaf to non-leaf. In a treemap, these comparisons are all rectangular area comparisons, but in the grouped bar chart condition comparing a non-leaf node to another node requires a more complicated cognitive process than the leaf-leaf comparison, which is a length comparison.

6.1 Methods

For each trial, we showed participants a chart with two highlighted nodes. In the grouped bar chart display, a highlighted node may be a group of bars. We then asked the participant to indicate which of the two was smaller (the discrimination task), and then what percentage the smaller was of the larger (the estimation task).

Before participants could accept our HITs, we required them to pass a qualification task that consisted of two example charts (one for each chart type) and two test questions (one for each chart type). The test questions were the same as the trial, but with multiple choice responses instead of free-text. For the estimation task answer choices, only one

was correct while the others were grossly incorrect, thus ensuring that participants understood the instructions.

NK: Insert statistics about how many took/passed qualification?

We tested 2 chart types (treemap and grouped bar chart), 5 data densities (256, 512, 1024, 2048, 4096 leaves), and 3 comparison types (leaf-leaf, leaf-non-leaf, non-leaf-non-leaf). A fully crossed design with 5 replications per cell resulted in $2 \cdot 5 \cdot 3 \cdot 5 = 150$ HITs per task. Each chart was sized at 600x400 pixels and displayed within a frame 600 pixels tall.

For each trial we recorded the time to completion for both the discrimination and estimation tasks, and participant's screen resolution, color depth, and browser type as reported by JavaScript. We used $N=24$ assignments per HIT.

NK: Insert statistics about:

- # of unique participants
- completion time (for all HITs)
- display resolution

Other notes. Did not ask structural questions because we were focusing on the value comparison task, which we hypothesize is the more common task.

6.1.1 Stimuli: Squarified Treemap

Discuss here? Why chosen?

6.1.2 Stimuli: Hierarchical Bar Chart

Fig. 6. A sample tree visualized as a (a) treemap, and (b) grouped bar chart.

There are many ways one could represent hierarchical data in a bar chart display. We designed our bar chart layout to use space as efficiently as possible while still maintaining some encoding of the structure of the data. Our central idea was to display each leaf as a bar, disregarding the level of that leaf in the hierarchy. Siblings that are leaves form a bar chart.

We first count the number of groups of leaves that share a parent (i.e., groups of siblings): this is the number of *cells* in the final display. Each cell contains a bar chart displaying the leaves in a group. Figure 6 shows an example tree visualized as a treemap and a grouped bar chart display. There are four groups of siblings in this example: one leaf is attached to the root and receives its own cell, while each of the other three groups of leaves are two levels from the root and each receive their own cell. The width of a bar indicates what level the leaf it encodes is at in the tree: the thicker the bar, the higher the level. In Figure ??, the top-left bar encodes the first-level leaf, so it is slightly thicker than the other bars. Finally, the color of the borders of the cell also indicate what level each bar is at: lighter borders indicate a higher level.

Every cell is the same size. We lay out the cells in a grid and attempt to minimize the aspect ratio of the cells while allowing enough width to show the densest cell (i.e., the cell with the most leaves). We therefore compute the minimum width required to display the densest cell, allowing the bars to shrink to a minimum width of one pixel while maintaining gaps of one pixel, and divide the available width by the minimum width to obtain the maximum number of columns possible. We then compute the number of rows that will minimize the aspect ratio of a cell given the maximum number of columns. This sometimes results in empty cells in the display.

NK: Discuss features of the chart that encode structure? Specifically, coloring, although all bars are the same color regardless of what level they are at.

6.1.3 Data generation

We create trees using a simple randomization process. Each tree has two levels below an implicit root node. We start at the root node and add a random number of children. To each of these children, we further add a random number of children. We then choose nodes in the tree at random (possibly including the root node) and add individual leaves until we reach the desired number of leaves.

- Creates ragged trees

6.2 Results

7 DESIGN RECOMMENDATIONS

8 CONCLUSION

ACKNOWLEDGMENTS

The authors wish to thank A, B, C. This work was supported in part by a grant from XYZ.

REFERENCES

- [1] M. Bertamini and D. Proffitt. Hierarchical motion organization in random dot configurations. *Journal of Experimental Psychology: Human Perception and Performance*, 26:1371–1386, 2000.
- [2] D. M. Bruls, C. Huizing, and J. J. van Wijk. Squarified treemaps. In *Data Visualization Eurographics/IEEE Symp.*, pages 33–42, 2000.
- [3] W. S. Cleveland and R. McGill. Graphical perception: Theory, experimentation, and application to the development of graphical methods. *J. Am. Statistical Assoc.*, 79:531–554, 1984.
- [4] W. R. Garner. *The processing of information and structure*. Erlbaum, 1974.
- [5] J. Heer and M. Bostock. Crowdsourcing graphical perception: Using mechanical turk to assess visualization design. In *ACM CHI*, 2010.
- [6] D. A. Keim. Designing pixel-oriented visualization techniques: Theory and applications. *IEEE Trans. on Visualization and Comp. Graphics*, 6:59–78, 2000.
- [7] A. M. MacEachren. *How Maps Work: Representation, Visualization, and Design*. Guilford, 1995.
- [8] B. Shneiderman. Tree visualization with tree-maps: 2-d space-filling approach. *ACM TOG*, 11:92–99, 1992.
- [9] D. Simkin and R. Hastie. An information-processing analysis of graph perception. *J. Am. Stat. Assoc.*, 82:454–465, Jun 1987.
- [10] I. Spence. The apparent and effective dimensionality of representations of objects. *Human Factors*, 46:738–747, 2004.
- [11] J. Stasko and E. Zhang. Focus+context display and navigation techniques for enhancing radial, space-filling hierarchy visualizations. In *IEEE Information Visualization*, pages 57–65, 2000.
- [12] S. S. Stevens. The psychophysics of sensory function. *American Scientist*, 48:226–253, 1960.
- [13] C. Ware. *Information Visualization: Perception for Design*. Morgan-Kaufmann, 2nd edition, 2004.
- [14] M. Wattenberg. Visualizing the stock market. In *ACM CHI Extended Abstracts*, pages 188–189, 1999.
- [15] D. Wigdor, C. Shen, C. Forlines, and R. Balakrishnan. Perception of elementary graphical elements in tabletop and multi-surface environments. In *ACM CHI*, pages 473–482, Apr 2007.
- [16] C. Ziemkiewicz and R. Kosara. The shaping of information by visual metaphors. *IEEE TVCG*, 14:1269–1276, Nov/Dec 2008.