Learning Perceptual Kernels for Visualization Design

Anonymous

Abstract—Visualization design benefits from careful consideration of perception, as different assignments of visual encoding variables such as color, shape and size affect how viewers interpret data. In this work, we introduce *perceptual kernels*: distance matrices derived from aggregate perceptual judgments. Perceptual kernels represent perceptual differences between and within visual variables in a reusable form that is directly applicable to visualization evaluation and automated design. We report results from crowdsourced experiments to estimate kernels for color, shape, size and combinations thereof. We analyze kernels estimated using five different judgment types — including Likert ratings among pairs, ordinal triplet comparisons, and manual spatial arrangement — and compare them to existing perceptual models. We derive recommendations for collecting perceptual similarities, and then demonstrate how the resulting kernels can be applied to automate visualization design decisions.

1 Introduction

Visual encoding decisions are central to visualization design. As viewers' interpretation of data may shift across encodings, it is important to understand how choices of visual encoding variables such as color, shape, size — and their combinations — affect graphical perception.

One way to evaluate these effects is to measure the perceived similarities (or conversely, distances) between visual variables. There are various ways of eliciting similarity judgments among visual variables. We broadly refer to subjective measures of judged similarity as *perceptual distances*. In this context, a *perceptual kernel* is the distance matrix of aggregated pairwise perceptual distances. These measures quantify the effects of alternative encodings and thereby help to create visualizations that better reflect structures in data. Figure 1a shows a perceptual kernel for a set of symbols; distances are visualized using grayscale values, with darker cells indicating higher similarity. The prominent clusters suggest that users will perceive similarities among shapes that may or may not mirror encoded data values.

Perceptual kernels can also benefit automated visualization design. Typically, automated design methods [27] leverage an *effectiveness* ranking of visual encoding variables with respect to data types (nominal, ordinal, quantitative). Once a visual variable is chosen, these methods provide little guidance on how to best pair data values with visual elements, instead relying on default palettes for variables such as color and shape. Perceptual kernels provide a means for computing optimized assignments to visual variables whose perceived differences are congruent with underlying distances among data points. In short, perceptual kernels enable the direct application of empirical perception data within visualization tools.

In this work, we contribute the results of crowdsourced experiments to estimate perceptual kernels for visual encoding variables of shape, size, color and combinations thereof. We compare a variety of judgment types: Likert ratings among pairs, ordinal triplet comparisons, and manual spatial arrangement. We also assess the resulting kernels via comparisons to existing perceptual models. We find that ordinal triplet matching judgments provide the most consistent results, albeit with higher time and money costs than pairwise ratings or spatial arrangement. We then demonstrate how perceptual kernels can be applied to improve visualization design through automatic palette optimization and by providing distances for *visual embedding* [8] of data points into visual spaces.

Manuscript received 31 March 2013; accepted 1 August 2013; posted online 13 October 2013; mailed on 4 October 2013.
For information on obtaining reprints of this article, please send e-mail to: tvcg@computer.org.

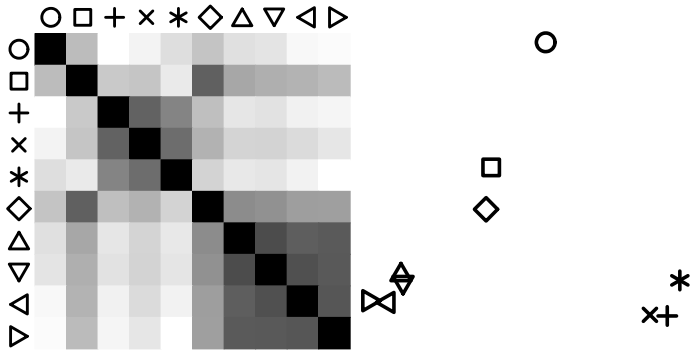


Fig. 1: (Left) A crowd-estimated perceptual kernel for a shape palette. The kernel was obtained using ordinal triplet matching. (Right) A two-dimensional projection of the palette shapes obtained via multidimensional scaling of the perceptual kernel.

2 Related Work

We draw on prior work in similarity judgments, interactions among perceptual dimensions, graphical perception and automated design.

2.1 Analysis of Perceptual Similarity Judgments

Prior research has analyzed similarity judgments to model perceptual spaces. Measurement methods involve asking subjects to rate or match multiple stimuli. One approach is to ask subjects to rate the perceived similarity of visual stimulus pairs using numbers on a specified numerical scale (such as a Likert scale). However, pairwise scaling can cognitively overload subjects and differences between subjects may confound analysis. These issues led to the use of simpler discrimination tasks involving ordinal judgments. Consider matching judgments over triplets: "Is A more similar to B than it is to C?" Ordinal judgments on triplets have been found more reliable and robust [20]. However, the number of pairs and triplets increases quadratically and cubically, respectively, with the number of visual stimuli. The method of spatial arrangement, where subjects rearrange stimuli in the plane such that their proximity is proportional to their similarity, was proposed as an efficient alternative [12]. In our experiments, we use direct judgment types, including Likert ratings among pairs, ordinal triplet rankings, and manual spatial arrangement.

Similarities may also be indirectly inferred from measurements such as subject response time (confusion time) or manual clustering [12]. For example, use of response time assumes that the similarity between two stimuli is related to the probability of confusing one with the other. Subjects are asked to quickly decide whether two given stimuli are the same; it is assumed that they take more time if the stimuli are more similar. In a clustering measure, subjects are asked to group given stimuli. It is assumed that the frequency with which two stimuli are placed in the same group is proportional to their similarity.

Embedding perceptual measurements in Euclidean space is an active line of research with impacts beyond psychology. Typically, such methods aim to model perceptual distances in terms of Euclidean distances. Torgerson's metric multidimensional scaling (MDS) [45] maps quantitative judgments onto Euclidean space. However, the use of triplet comparisons requires one to map ordinal judgments. Shepard and Kruskal [22, 32] proposed non-metric multidimensional scaling (NMDS) to handle general cases of perceptual measurements. Their formulation requires a complete ranking of all stimulus pairs, prompting more general formulations of NMDS that derive perceptual distances from only a partial set of ordinal judgments [1, 31, 47]. These methods allow distances to be inferred from only a subset of all possible comparisons. Tamuz et al. [44] further propose an adaptive sampling method for more efficient learning of crowdsourced kernels.

2.2 Dimensional Integrality of Perceptual Dimensions

Visual variables are often applied in tandem to represent multidimensional data. How does perception of one visual variable change when combined with another? To address this question, researchers have investigated interactions between perceptual dimensions [2, 11, 33, 35]. These investigations led Garner and Felfoldy [11] to introduce a distinction between two types of stimulus dimensions:*integral* and *separable*. Visual stimulus dimensions are considered integral if they interfere with or facilitate perception of the other. Dimensions are considered separable if they do not. For integral dimensions, redundant encoding (representing the same data with multiple visual variables) can improve task performance. When the dimensions are fully separable, redundant encoding does not affect task performance. If a task requires selective attention (focusing on one dimension while filtering out the other) integral dimensions can interfere, impairing task performance. Integrality and separability do not form a crisp dichotomy, but rather a continuum with varying degrees of interaction [11].

Integrality can also be measured in terms of the structure of perceptual spaces. Prior research [2, 33, 45] provides some evidence that, for integral dimensions, perceptual distances over multiple visual variables form a Euclidean (L_2) metric. For separable dimensions, they form a city-block (L_1) metric. For example, Attneave [2] found that a city-block metric better explained his experimental measurements than a Euclidean metric for size and shape and for size and brightness. Torgerson [45] showed that color value and chroma elicit judgments consistent with a Euclidean metric. We revisit these findings in our analysis of crowd-estimated perceptual kernels.

The importance of this dichotomy from the perspective of perceptual kernels is that it may give hints about how to build new perceptual kernels for multidimensional visual stimuli by using already-known perceptual distances of individual dimensions.

2.3 Graphical Perception

A related area of research is *graphical perception* [6]: the decoding of data presented in graphs. How do choices of visual variables such as position, size, shape or color impact visualization effectiveness? Bertin was among the first to systematically study visual variables' "capacities for portraying given types of information" [3]. Following Bertin, researchers in multiple disciplines have conducted human subjects experiments [6, 14, 21, 24, 37, 38, 46] and proposed perceptually-motivated rankings of visual variables for nominal, ordinal or quantitative data [6, 24, 25, 27, 36]. Researchers have also investigated how different choices of design parameters such as aspect ratio [5, 13, 42], chart size [16, 23], axis labeling [43] and animation design [17, 29] influence the effectiveness of graphs. This work typically compares the effectiveness of alternative visual variables. In contrast, perceptual kernels enable analysis of visual encoding assignments both *within* and *between* specific classes of visual encoding variables.

2.4 Automated Visualization Design

Mackinlay's [27] Automatic Presentation Tool (APT) is one of the most influential systems for automated visualization design. Mackinlay formulates visualizations as sentences in a graphical language and argues that good visualizations are those that meet his criteria of *expressiveness* and *effectiveness*. According to Mackinlay, a visualization is expressive if it faithfully presents the data, without implying false inferences. A visualization is effective if the chosen visual variables are accurately decoded by viewers. APT employs a composition algebra over a basis set of graphical primitives derived from Bertin's encodings to generate visualizations. The system then selects the visualization that best satisfies formal expressiveness and effectiveness criteria. To operationalize effectiveness, APT uses a rank ordering of visual variables by data type, which is informed by prior studies in graphical perception (e.g., [6, 34]).

APT does not explicitly take user tasks or interaction into account. To this end, Roth et al. [30] extend Mackinlay's work with new types of interactive presentations. Similarly, Casner [4] builds on APT by incorporating user tasks to guide visualization generation. Some of these ideas are now used for visualization recommendation within Tableau, a commercial visualization tool [26].

Demiralp et al. [8] propose *visual embedding* as a model for visualization construction and evaluation. A visual embedding is a function from data points to a space of visual primitives that measurably preserves structures in the data (domain) within the mapped perceptual space (range). This framework can be used to generate and evaluate visualizations based on both underlying data and — through the choice of preserved structure — desired perceptual tasks. To assess structural preservation, the *visual embedding* model requires perceptual distance measures for a given visual embedding space. In some cases, existing perceptual spaces, such as CIELAB color space, can be used to perform embeddings [7]. In this work, we evaluate crowdsourcing methods to estimate perceptual kernels for visual encoding variables that lack suitable models. The resulting kernels can be applied directly in visual embedding procedures or used to derive and evaluate more general perceptual models.

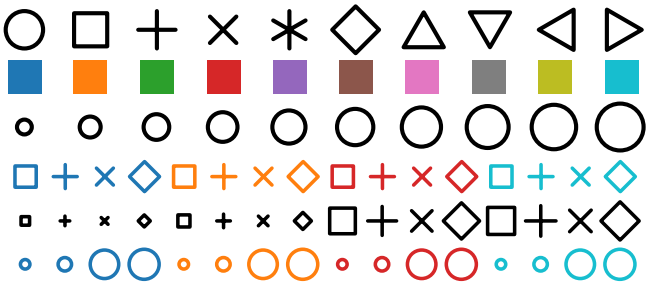


Fig. 2: Palettes of visual stimuli used in our experiments: color, shape, size, shape-color, shape-size, size-color.

3 Research Goals and Experiment Overview

Our ultimate goal in introducing perceptual kernels is to facilitate automated visualization design. In order to do so, we must be able to estimate perceptual kernels reliably. Our first research goal was to evaluate and compare multiple approaches for collecting crowdsourced judgments to construct perceptual kernels. Our second research goal was to demonstrate the utility of these kernels for generating and evaluating both visualizations and new perceptual models.

We conducted two experiments to learn perceptual kernels for visual encoding variables of shape, color, size and their combinations. The first experiment elicited judgments for univariate encodings, the second for bivariate encodings. The two experiments share the same procedure: collect similarity judgments under various rating schemes, construct perceptual kernels, then analyze the results.

3.1 Visual Stimuli

We used color and shape stimuli from palettes in Tableau, a commercial visualization tool. Tableau's shape and color palettes were manually designed with consideration of perceptual issues such as discriminability, saliency and naming of colors [40], and robustness to spatial overlap of shapes. As such, these palettes constitute a good base from which to evaluate perceptual kernels. Also, using palettes from a popular visualization tool provides ecological validity for our study. Both the basic color and shape palettes have ten distinct values. For size, we

used ten circles with linearly increasing area. We obtained perceptual kernels for each of these stimulus sets and their bivariate combinations. In total, we evaluated the six palettes shown in Figure 2: *color, shape, size, shape-color, size-color, shape-size*.

3.2 Judgment Types

We compared five similarity judgment types, each differing in terms of elicitation strategy or reported precision:

Pairwise rating on 5-Point Scale (L5): Subjects were sequentially presented pairs of visual stimuli and asked to rate the similarity of each pair on a 5-point Likert scale (Figure 3). The order between and within pairs was randomized for each subject. Task progress was visualized as an upper-triangular matrix, which was filled in as the subject provided ratings. This representation allowed subjects to see all their ratings together and readjust them as needed. Once all pairwise ratings were completed, subjects could click any cell and change the rating for the corresponding pair. The design goal was to help subjects distribute their ratings within the Likert scale so that the most different stimulus pairs get the highest rating while the most similar, non-identical stimulus pairs get the lowest possible rating. This also helps mitigate the effects due to differences between internal scales of subjects, a well-known problem for subjective pairwise scaling [20].

Pairwise rating on 9-Point scale (L9): Same as the task above, except that a 9-point Likert scale was used.

Triplet ranking with matching (Tm): Subjects were sequentially presented triplets of stimuli, with one indicated to be a reference. We asked subjects to decide which of the other two stimuli was the most similar to the reference (Figure 4). The order between and within triplets was randomized for each subject.

Triplet ranking with discrimination (Td): Subjects were sequentially presented triplets of stimuli and asked to decide which one was the most dissimilar to the other two (Figure 5). The order between and within triplets was randomized for each subject.

Spatial arrangement (SA): Subjects were asked to manually arrange stimuli in the plane such that the 2D distances between pairs are proportional to their perceived dissimilarity (Figure 6). The initial layout was randomized for each subject. To standardize interpretation of the instructions, we provided an example demonstrating the continuous nature of the judgments.

3.3 Experimental Platform & Subjects

We collected similarity judgments by submitting jobs to Amazon's Mechanical Turk (MTurk), a popular micro-task market that is regularly used for online human subjects experiments. For example, Heer & Bostock [14] reproduced prior laboratory experiments on spatial encoding [6] and alpha contrast [41], demonstrating the viability of crowdsourced graphical perception studies. We ran thirty separate (six visual variables $\times$ five judgment types) MTurk jobs. Each job was completed by 20 Turkers, for a total of 600 distinct subjects. We limited the participant pool to Turkers based in the United States with a minimum 95% approval rate and at least 100 approved tasks.

3.4 Procedure

For all but the spatial arrangement (SA) task, subjects carried out the experiments in five steps. Subjects were first presented a description of the task with an option of accepting it. Once the task was accepted, subjects completed a training session using an interface identical to the actual task interface but populated with different visual stimuli. After the training session, subjects were prompted with the full set of visual stimuli and asked to think about the most similar and dissimilar stimuli in the set (Figure 7). Once they were ready, subjects completed the experimental task. In the last step, they provided comments on their rating or ranking strategies and submitted their results.

The SA experiments were carried out in two simple steps; The task interface and instructions were directly presented to subjects upon introduction. Instructions included a spatial arrangement example (Figure 6). Once the subjects were satisfied with the layout, they provided comments on their strategies and submitted their layout.

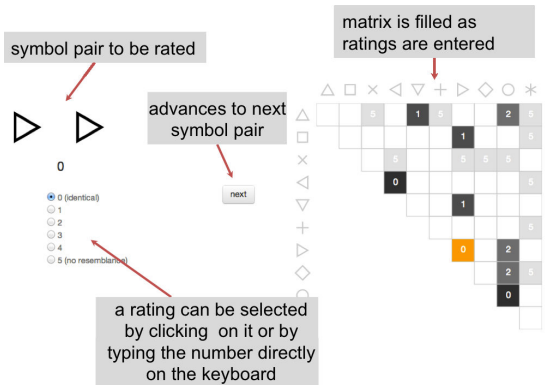


Fig. 3: Experiment interface for the pairwise rating task on a Likert scale of five (L5).

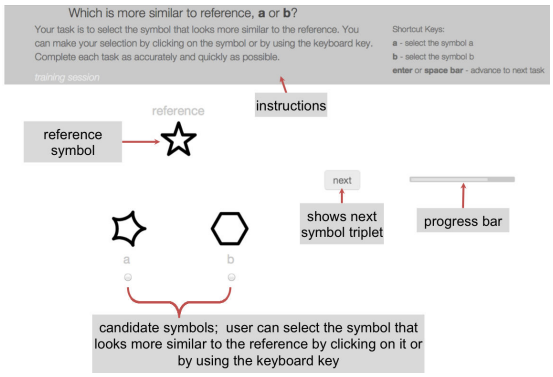


Fig. 4: Interface for the triplet matching task (Tm).

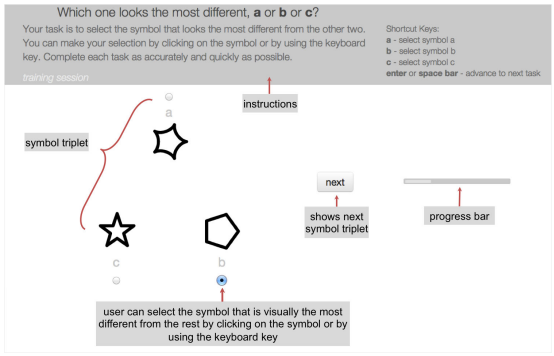


Fig. 5: Interface for the triplet discrimination task (Td).

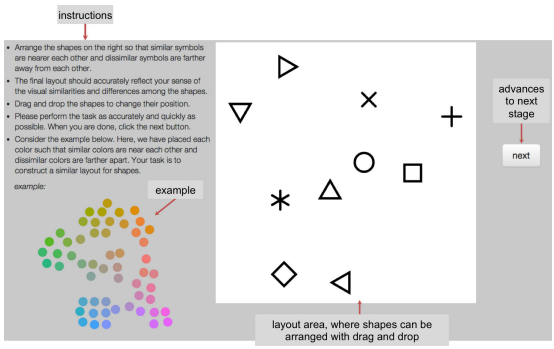


Fig. 6: Interface for the spatial arrangement task (SA). Subjects can rearrange visual stimuli (here shapes) with drag and drop.

Fig. 7: Visual stimuli overview. We asked subjects to consider and compare the stimuli before starting the experimental task.

3.5 Data Processing

Our pairwise judgment tasks directly produce a distance matrix among visual stimuli; we simply rescale the per-user ratings to the range [0,1]. For triplet judgments, we derive per-user kernels from a set of rank-ordered triplets using generalized non-metric multidimensional scaling [1]. In both cases, we then average the per-user kernels and re-normalize the result to form an aggregate perceptual kernel.

To safeguard data quality, we use errant ratings of identical stimuli pairs (both in the pairwise and triplet cases) to filter "spammers." In the pairwise cases, subjects were instructed to rate the similarity of identical pairs as 0. They were also expected to match or filter identical stimuli pairs in the triplet cases. We excluded the data from subjects who failed in 40% or more of these judgments.

For spatial arrangements, we align each arrangement with every other arrangement using similarity transforms via Procrustes analysis [19]. We designate the arrangement that requires the minimum total transformation to align with others as the reference arrangement. We then align all responses to this reference arrangement, use in-plane Euclidean distances to construct distance matrices for each subject, and then normalize the results. To combat spamming, we removed layouts whose alignment error was greater than a threshold of two standard deviations away from the mean alignment error. Finally, we average the distance matrices and normalize the result to obtain a perceptual kernel.

Throughout the paper, we present the resulting perceptual kernels as matrix diagrams alongside a 2D projection obtained using multidimensional scaling. These projections are intended to provide a more intuitive, overall sense of the kernel. Note, however, that each projection is a lossy representation, in some cases providing only partial insight into the kernel structure.

4 Experiment 1: Univariate Kernels

In the first experiment, we estimated perceptual kernels for stimuli that change only in one perceptual dimension (i.e., univariate visual variables). We chose the visual variables shape, color, and size due to their common use in practice. For values of shape and color, we used Tableau's default shape and color palettes, each of which has ten values. We presented colors to subjects as rectangular chips, which is customary in perceptual experiments. For the size variable, we used ten circles with linearly increasing area.

4.1 Estimated Univariate Perceptual Kernels

Figure 10 visualizes the resulting kernels for each palette and judgment type. We summarize specific results for each palette below.

4.1.1 Shape

Figure 1 shows a matrix and two-dimensional MDS projection of the perceptual kernel estimated from distinct triplet (Td) judgments. The MDS projection shows distinct perceptual shape clusters. Across all kernels (Figure 10), we see strong groupings among triangles and stroked shapes, and a looser cluster of other filled shapes.

4.1.2 Color

Figure 8 shows the matrix and two-dimensional MDS projection for color values. From the MDS projection we readily see that subjects judged color similarity primarily by hue and secondarily by lightness.

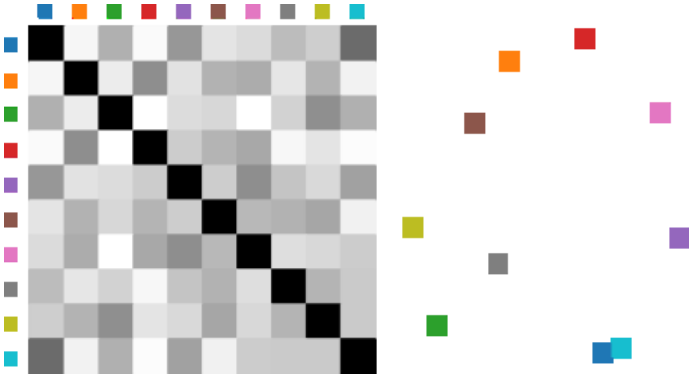
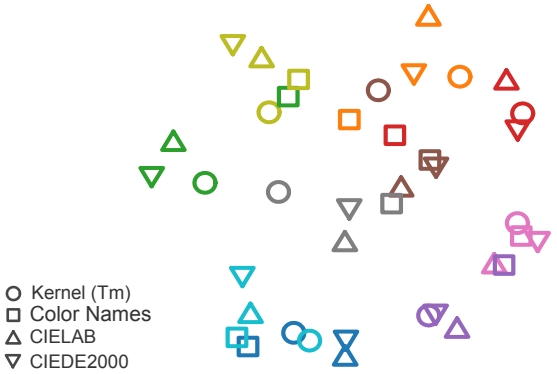


Fig. 8: (a) A crowd-estimated perceptual kernel for the color palette. The kernel was obtained using the triplet matching (Tm) task. (b) A two-dimensional projection of palette colors obtained via multidimensional scaling of the perceptual kernel.



	Kernel (Tm)	CIELAB	CIEDE2000	Color Names
Kernel (Tm)	1.00	0.67	0.59	0.75
CIELAB	0.67	1.00	0.87	0.81
CIEDE2000	0.59	0.87	1.00	0.77
Color Names	0.75	0.81	0.77	1.00

Fig. 9: (Top) Projections of a crowd-estimated color kernel and kernels induced by CIELAB, CIEDE2000 and color name distances, all aligned by similarity transforms. Plotting symbols were chosen through visual embedding of the rank correlations between metrics. (Bottom) The rank correlation between kernels.

To further validate the crowd-estimated kernels, we can compare them to kernels derived from existing color models. CIELAB is an approximately perceptually uniform color space with a lightness component L* and opponent color components a* and b*. CIEDE2000 is a more complex color difference formula that was developed to better fit empirical perceptual judgments than Euclidean LAB distances. Heer and Stone [18] introduced distances based on color-name associations to reflect linguistic boundaries among colors. Here, we use the Hellinger distance between multinomial color name probability distributions estimated from the XKCD color naming survey [28].

Figure 9 compares the triplet matching (Tm) kernel with kernels constructed using CIELAB, CIEDE2000 and color name distance [18] distance measures. All kernels are strongly correlated, but we also see some variation, consistent with the fact that longer distances in existing perceptual color spaces tend to be less accurate than short proximal judgments. Interestingly, of the existing models color name distance correlates most highly with the crowd-estimated kernel. We hypothesize that perceptual judgments from crowd participants are influenced by color name associations in addition to lower-level features.

4.1.3 Size

As shown in Figure 10, of the three visual variables we considered, size is the most robust across judgment task types. The MDS projec-

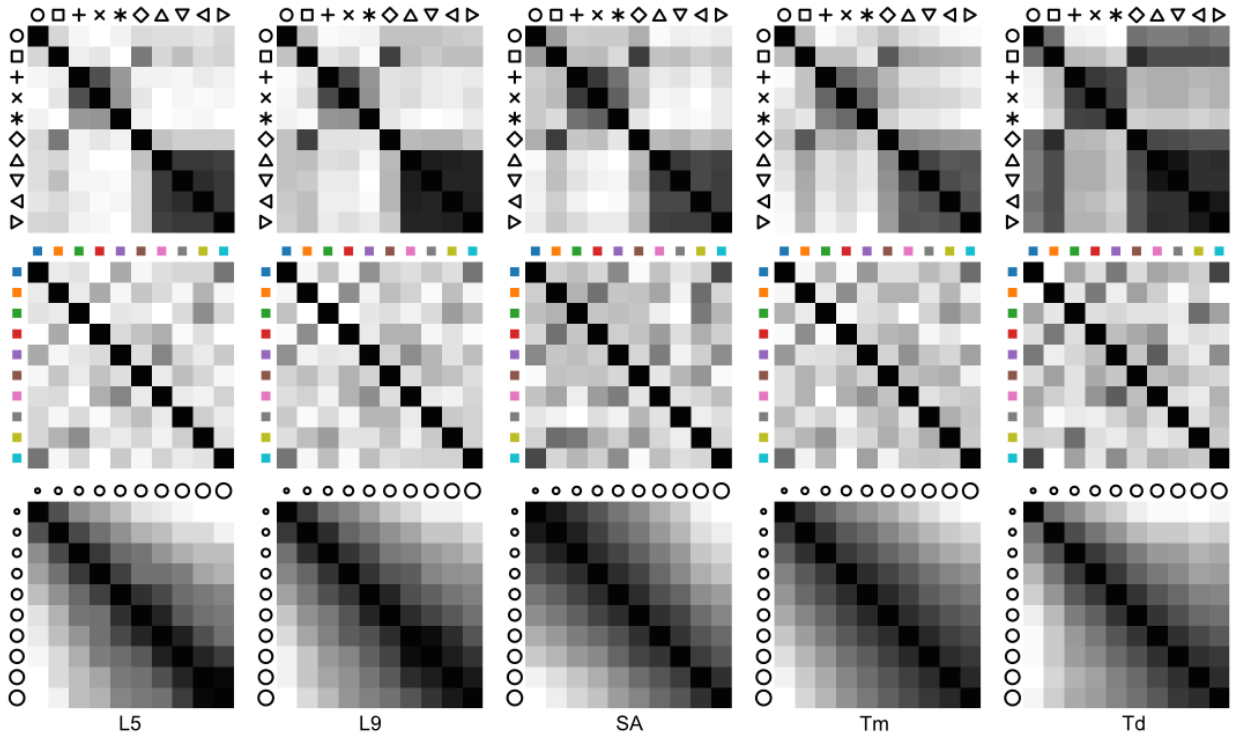


Fig. 10: Experiment 1 Results. Univariate perceptual kernels for shape, color and size palettes across different judgment types. Darker colors indicate higher perceptual similarity. For each palette, the matrices exhibit consistent structures across judgment types.

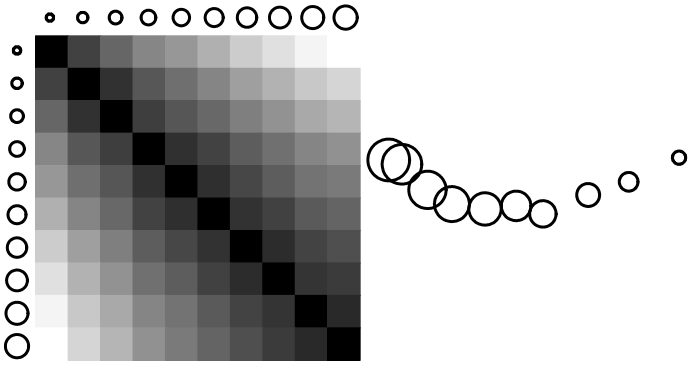


Fig. 11: (a) A crowd-estimated perceptual kernel for the size palette. (b) A two-dimensional projection of size values obtained via multidimensional scaling of the perceptual kernel.

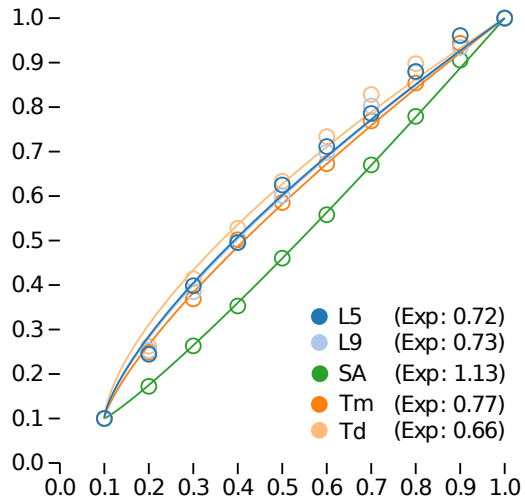


Fig. 12: Stevens' Power Law fits to kernel-estimated area magnitudes.

tion in Figure 11 clearly demonstrates a one-dimensional structure, in which linear increases in area map to non-linear perceptual distances. Non-linearity of area judgments is consistent with perceptual models such as the Weber-Fechner Law [10] and Stevens' Power Law [39]. Stevens posits a power-law relationship between the magnitude of a physical stimulus and its perceived intensity: $S \sim I^{\beta}$, where S and I are the sensed and the true intensities, respectively.

Figure 12 shows Stevens' Power Law fits and corresponding exponent values for each judgment type. Pairwise and triplet kernels result in exponents consistent with the literature on area estimation (0.7-0.8). For spatial arrangement (SA) we find an exponent larger than one, which is inconsistent with prior work. To compute these fits, we calculated individual area estimates from each row of the kernel, treating the diagonal value as a reference. We then averaged the resulting magnitude estimates and directly perform least-squares fitting of the power law exponent. We constrain the lowest and highest areas to their true values, as the full palette was known to subjects from the outset. However, the resulting exponents are robust across such modeling decisions.

5 Experiment 2: Bivariate Kernels

In the second experiment, we estimated perceptual kernels for stimuli that change in two perceptual dimensions (i.e., bivariate visual variables). We chose four elements from each of the univariate palettes and used their pairwise combinations to create three bivariate palettes with 16 values: *shape-color*, *size-color*, and *shape-size* (Figure 2). To test interactions among perceptual dimensions, we intentionally included both highly similar and highly dissimilar values from the univariate palettes (e.g., two small sizes and two large sizes).

We did not use the complete set of elements from the univariate palettes, as this would cause the bivariate palettes to become too large to practically run our experiments. A bivariate variable with 100 values requires rating 4,950 ($=100 \times 99/2$) pairs. As discussed previously, this number is even larger when using triplet ratings.

5.1 Estimated Bivariate Perceptual Kernels

Figure 13 visualizes the estimated bivariate kernels for each palette and judgment type. Figure 14 shows both kernels and two-dimensional

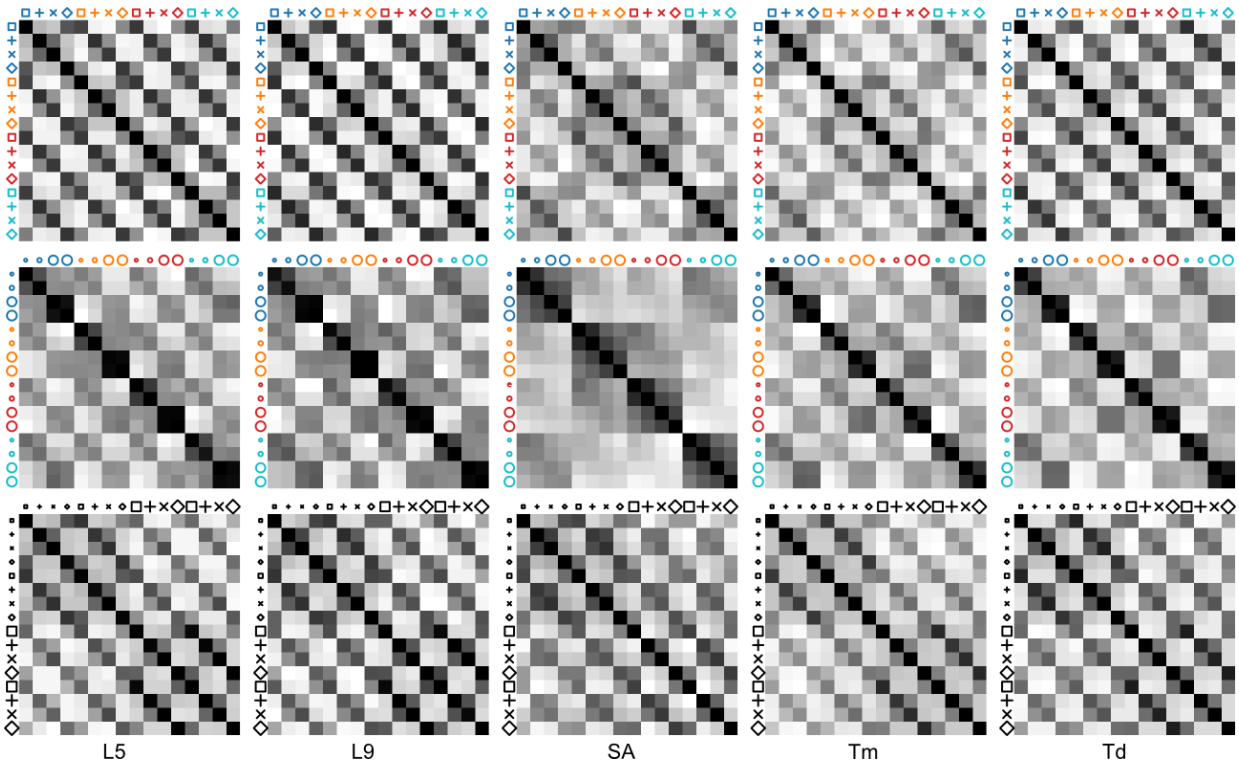


Fig. 13: Experiment 2 Results. Bivariate perceptual kernels for the shape-color, shape-size, and size-color palettes across judgment types. Darker colors indicate higher perceptual similarity.

MDS plots for triplet matching (Tm) judgments. In most cases we observe balancing among visual variables: large distances in one variable dominates smaller distances in the other. We also note limitations of the MDS plots in Figure 14: the 2D projection collapses smaller distances, resulting in overlapping points. The actual structure is better described by three dimensions, in which these clusters are more distributed. We summarize specific results for each palette below.

5.1.1 Shape-Color

For all kernels but triplet discrimination (Td), shape-color stimuli form four dominant intersecting clusters, grouped by the most similar color and shape values. For the Td kernel shape dominates color entirely, forming four clusters of distinct shapes and mixed colors. As we will describe in the next section, this is likely due to the failure of triple discrimination to elicit more fine-grained comparisons.

5.1.2 Shape-Size

Across all judgment types we see results similar to the size-color kernels: the shape-size kernels exhibit four dominant clusters, grouped by the most similar shape and size values.

5.1.3 Size-Color

Across all judgment types, the size-color kernels exhibit four dominant clusters, grouped by the most similar size and color values. Three-dimensional MDS plots (see supplementary material) reveal additional stratification by color value.

5.2 Analysis of Dimensional Integrality

An important issue with visual variables is their potential interactions with each other when used to encode multiple dimensions of data. Our bivariate palettes are examples of two-dimensional stimuli. Prior research states that dimensions of a visual stimulus are *separable* if they do not confound or facilitate perception of the other and are considered *integral* if they do [11].

Researchers further argue (e.g., [2, 45, 33]) that if the dimensions constituting a multidimensional stimulus are integral then the multidimensional perceptual distances can be approximated using a Euclidean (L_2) metric. If the dimensions are separable, then the distance in the multidimensional stimulus space can be better approximated with the city-block (L_1) metric.

To assess if either of these metric structures holds for estimated perceptual kernels, we fit the following weighted power model to predict the values of the bivariate shape-color, shape-size, and size-color kernels based on the corresponding univariate kernels:

$$d_{ij} \sim b_0 + ((b_1 d_1)^n + (b_2 d_2)^n)^{1/n}$$

Here, d_{ij} is the observed perceptual distance between two bivariate stimuli i and j. d_1 is the univariate distance between i and j on the first perceptual dimension and d_2 is the univariate perceptual distance on the second dimension. b_1 and b_2 are scaling parameters acting on the perceptual space, which account for any non-uniformity in the strength of the individual perceptual dimensions. Prior work [11] suggests that the value of n depends on the level of integrality between dimensions. A value of $n = 1$ would indicate total separability, whereas a value of $n = 2$ would indicate complete integrality. We fit the weighted power model to our experimental data using non-linear regression routines in Matlab. We set $b_2 = 1 - b_1$ without constraining the sum to be 1.

	shape-color				shape-size				size-color			
	b_0	b_1	n	llik	b_0	b_1	n	llik	b_0	b_1	n	llik
L5	0.05	0.78	1.04	186	0.12	0.72	1.24	178	0.10	0.52	1.28	168
L9	0.10	0.86	0.99	198	0.13	0.77	1.12	181	0.08	0.54	1.13	169
SA	0.22	0.56	1.18	191	0.21	0.78	1.46	144	0.18	0.28	1.48	131
Tm	0.25	0.65	1.27	239	0.24	0.70	1.24	214	0.20	0.56	1.55	209
Td	0.24	0.89	1.07	222	0.23	0.84	1.10	189	0.19	0.54	1.45	166

Table 1: Estimated parameters of the weighted power model fitted to perceptual kernels. $b0$ is the intercept (or bias), b_1 is the scaling of the first dimension, $b_2 = 1 - b_1$ is the scaling factor for the second dimension, and n is the exponent of the model. Across palettes, triplet matching (Tm) provides the best prediction (highest log-likelihood) of bivariate distances from univariate kernels.

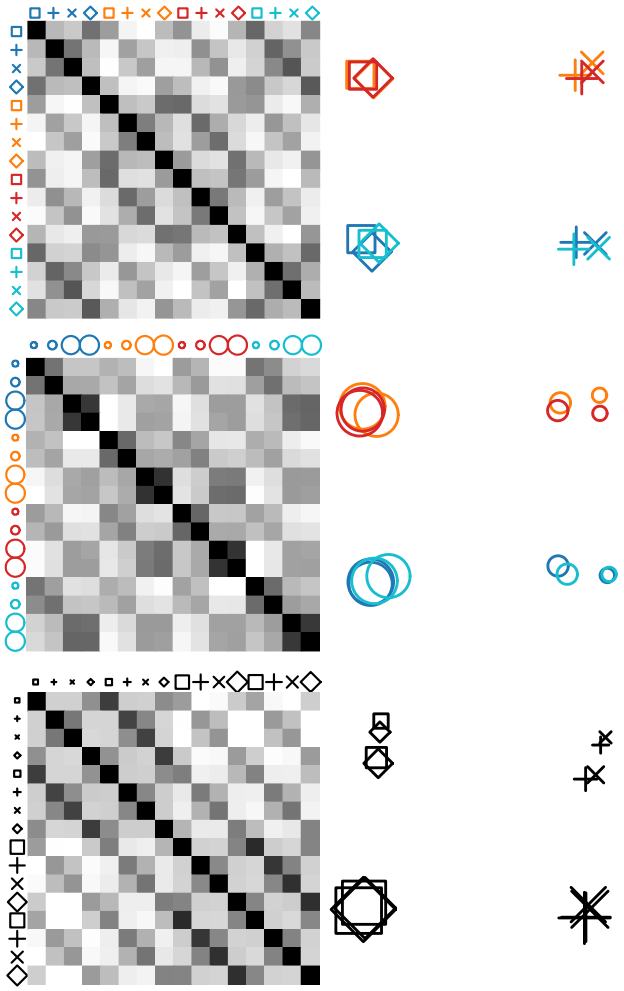


Fig. 14: (Left) Crowd-estimated kernels (Tm) for shape-color, size-color and shape-size palettes. (Right) Two-dimensional projections of the kernels obtained by multidimensional scaling.

Table 1 summarizes the fit model parameters and the goodness-of-fit in terms of log-likelihood values. With the exception of spatial arrangement (SA), each judgment type exhibits similar values of the scaling parameters b_1 and b_2 (=$1-b_1$), indicating the degree by which each dimension is scaled. In accordance with the prior literature, some level of integrality (n values intermediate between 1 and 2) is seem across all variables, but more so on average for interactions involving size (particularly size and color) than for color and shape. As indicated by the model log-likelihoods, across all palettes triplet matching judgments provide the most accurate prediction of bivariate distances from univariate kernels.

6 Comparison of Judgment Tasks

One goal of this work is to understand the trade-offs among different judgment tasks. In addition to the perceptual analyses in previous sections, we performed comparative analyses considering factors such as collection cost, agreement and robustness. We then provide recommendations based on the results of our analysis.

6.1 Variance and Cost

Table 2 presents summary statistics for each judgment type. Across judgments, triplet matching (Tm) exhibits the lowest cross-subject variance and lowest unit task time. The low per-task time is consistent with the binary perceptual judgment requires. Other tasks require considering more potential responses: three in the case of triplet discrimination, and either five or nine for pairwise Likert ratings. Unsurprisingly, L9 exhibits the longest per-judgment time. However, pairwise rating requires fewer total judgments, leading to lower overall experiment time and cost than triplet comparisons. Spatial arrangement (SA) is by far the fastest, and hence cheapest, elicitation method.

	univariate					bivariate				
	σ_m	μ_t	σ_t	μ_T	\$	σ_m	μ_t	σ_t	μ_T	\$
L5	0.03	3.29	2.25	180.96	0.75	0.04	3.28	3.02	446.67	2.00
L9	0.04	3.63	2.22	199.74	0.75	0.05	3.58	3.24	486.79	2.00
SA	0.04			43.18	0.20	0.03			180.79	0.35
Tm	0.02	2.51	2.42	345.73	1.00	0.01	2.36	2.11	1401.48	3.50
Td	0.02	3.18	2.48	439.25	1.00	0.03	2.37	1.81	1407.21	3.50

Table 2: Summary comparison of judgment task types; standard deviation across per-subject kernel distances (σ_m), average judgment time (μ_t), standard deviation of average judgment time (σ_t), the average duration of the experiment (μ_T), and per Turker compensation (\$). All time measurements listed are in seconds. Measurements μ_t and σ_t are not directly applicable to SA, and so left blank.

	shape	color	size	shape-color	shape-size	size-color	**Avg.**
L5	0.87	0.80	0.96	0.91	0.93	0.86	0.89
L9	0.87	0.78	0.96	0.91	0.93	0.87	0.89
SA	0.78	0.58	0.89	0.86	0.90	0.62	0.77
Tm	0.84	0.79	0.97	0.91	0.94	0.86	0.89
Td	0.85	0.75	0.94	0.87	0.94	0.86	0.87
Avg.	0.84	0.74	0.94	0.89	0.93	0.81	0.86

Table 3: Average rank correlations between each estimated kernel and all other perceptual kernels for the same palette.

6.2 Correlations

To better understand the degree of compatibility between the five judgment tasks, we compared their corresponding perceptual kernels. To quantify the degree of similarity between perceptual kernels, we use Spearman's rank correlation coefficient. While we believe rank correlation is the most appropriate measure, we note that standard correlation coefficients (Pearson's product moment) provide similar results.

Table 3 summarizes these correlations. SA has the lowest average correlation across all variables; the other task types exhibit similar correlations. We see that both task type and visual variable affect the level of correlation. Color has the least agreement while size has the most, suggesting a potential relationship between the dimensionality of the underlying perceptual space and agreement across task types. When the perceptual space has low dimensionality, tasks may become easier, reducing the effects of cognitive load and higher degrees of freedom.

6.3 Sensitivity

How sensitive are the kernels to the number of subjects who participate? To address this question, we ran a sensitivity analysis on judgment tasks across univariate and bivariate kernels (Figure 15). We randomly remove subjects from the experiments and compare the original perceptual kernels with those derived from the reduced datasets.

The results show that on average spatial arrangement (SA) is the least robust to changes in data size, while triplet matching (Tm) is the most robust. The sensitivity to subject pool size is also affected by the visual variable used. All judgment types are are highly stable with the size variable, as it forms a relatively simple perceptual space. Conversely, estimated kernels are less stable with color and, to a lesser degree, with shape. The five judgment types are less stable with the univariate variables than they are with the bivariate variables, though this is likely due (at least in part) to the very specific stimuli chosen for our bivariate experiments. Overall, all the judgment types are considerably robust. Even in the case of SA, the rank correlation is above 0.6 when 80% of the experimental data is removed.

6.4 Discussion: Which Judgment Task to Use?

Our analyses have identified trade-offs among judgment types. Which should be preferred? We now consider each class of judgments in turn.

Spatial Arrangment (SA). Spatial arrangement is clearly the fastest

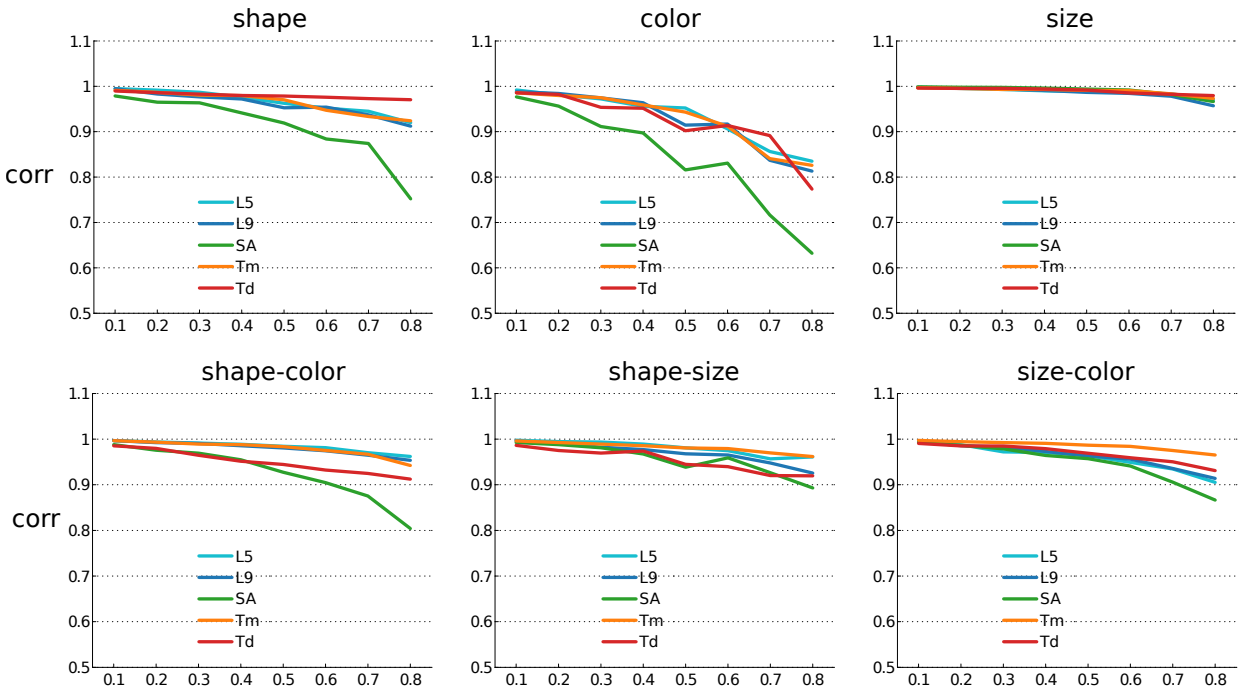


Fig. 15: Sensitivity of judgment types to the removal of subject data. The x-axis indicates the percentage of subjects dropped from each experiment; the y-axis indicated rank correlation. All kernels are highly stable for the size palette, as it is relatively simple perceptual space. For shape and color, the stability decreases faster, with the SA task deviating considerably from the others.

and cheapest method for eliciting perceptual kernels. However, for all other measures it is the worst-performing judgment task among the five considered. We believe there are multiple reasons for this outcome. First, SA tasks are the least structured, leading to higher variance across subjects. Second, by design SA tasks are inherently limited to two-dimensional structures. Unlike the other judgment tasks, SA can not accurately express higher dimensional structures. This limitation is especially problematic for the case of color (which is known to be best modeled using three dimensions) and for judgments spanning multiple perceptual dimensions.

Pairwise Likert Ratings (L5 & L9). Pairwise rating fared admirably in our experiments. These ratings are faster and cheaper to elicit than exhaustive triplet comparisons. However, triplet matching (Tm) exhibits lower variance and slightly improved robustness. One potential issue with Likert judgments is a possible confound of scale cardinality. When the number of stimuli outnumber the Likert scale levels (in this case, 5 or 9), judgments are limited in their precision, as certain fine-grained differences may be inexpressible. That said, we do not see any clear evidence of this issue affecting the results of this work. One potential explanation is that such high-precision judgments, while desirable in theory, are in fact dominated by between-subject variation.

Triplet Comparison (Tm & Td). Setting aside issues of experiment time and cost, our analyses indicate that triplet matching with a reference (Tm) is the preferred judgment type. Triplet matching exhibits the lowest variance in estimates, is the most robust across the number of subjects, and results in the most accurate prediction of bivariate kernels from univariate inputs. Triplet matching also involves the shortest unit task time (as opposed to overall experiment duration). Triplet matching involves a two-alternative forced-choice, and so arguably is the simplest and most "perceptual" of the tasks considered.

Why does triplet matching (Tm) outperform triplet discrimination (Td)? First, as noted above, it involves a simpler binary (as opposed to trinary) decision. Second, triplet matching elicits more fine-grained distinctions. Consider three stimuli A, B and C, and assume the "true" distances are as follows: $d(A,B) = 0.1$, $d(A,C) = 0.8$, $d(B,C) = 0.9$. In the case of Td, when subjects see the triplet A, B, C they should pick the most distinct, which in this case is C. In the case of Tm, some judgments will use C as the reference. Subjects are then forced to choose either A or B as the most similar. In this case, most subjects will probably pick A (as $0.8 < 0.9$). Thus triplet matching encourages more fine-grained distinctions, providing more information for the subsequent scaling. This comes with the potential cost of requiring multiple judgments per triplet, using different references. However, in our experiments we use the same total number of judgments as triplet discrimation and still see better, more robust results.

As a result of these considerations, we advocate for the use of triplet matching (Tm) judgments unless prohibited by time or cost. There are also various means of scaling triplet judgments to larger palettes. One method is to subdivide the stimulus set and parcel out different subsets to different subjects. A complementary method is to use adaptive sampling methods [44] for more scalable, active learning of perceptual kernels. We defer further exploration of these options to future work.

7 Applications

In this section, we present example applications using perceptual kernels for automated visualization design. In the first application, we generate re-orderings of the Tableau palettes to optimize perceptual discriminability. In the second application, we demonstrate how perceptual distances provided by the kernels can be used to perform visual embedding for optimized assignment of palette entries to data points.

7.1 Automatically Designing New Palettes

Given an estimated perceptual kernel, we can use it to revisit existing palettes. For example, we can choose a set of stimuli that maximizes perceptual distance or conversely minimizes perceptual similarity according to the kernel. Figure 16 shows the n most discriminable subsets of the shape, size, and color variables. (We include size for completeness, though in practice this palette is better suited to quantitative, rather than categorical, data.) To compute a subset with n elements, we first initialize the set with the variable pair that has the highest perceptual distance. We then add new elements to this set, by finding the variable whose minimum distance to the existing subset is the maximum (i.e., the Hausdorff distance between two point sets).

Note that there are several ways we might re-order palettes. For example, we could perform a global optimization for each value of n. However, one advantage of the method used here is that it is stable: a given subset palette grows only by adding new elements, without replacing the existing ones. We do not need to change the visual variables already assigned if new data values are added.

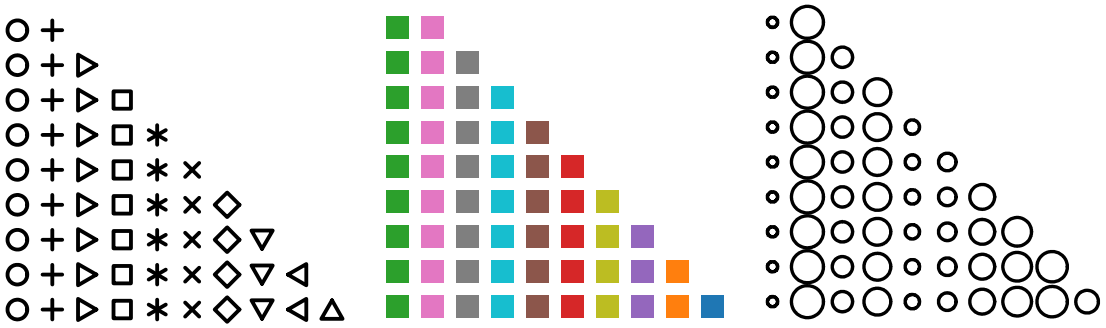


Fig. 16: Shape, color and size palettes re-ordered to maximize perceptual discriminability according to triplet matching (Tm) kernels.

It is instructive to compare the re-ordered palettes with the two-dimensional MDS projections of the kernels. For example, the new four-element shape palette contains a representative from each of the four clusters seen in Figure 1b. Our shape and color palettes have been re-ordered such that more perceptually discriminable stimuli are used first. Of course, this example considers only perceptual aspects, assuming equivalent distances among input data values.

7.2 Visual Embedding

Perceptual kernels can also guide *visual embedding* [8] to choose encodings that preserve data-space distance metrics in terms of kernel-defined perceptual distances. To perform discrete embeddings, we find the optimal distance-preserving assignment of palette items to data points (e.g., using simulated annealing or other optimization methods).

The scatter plot in Figure 9 compares color distance measures. The plotting symbols were chosen automatically using visual embedding. We use the correlation matrix between color models as the distances in the data domain, and the triplet matching (Tm) kernel for the shape palette as the distances in the perceptual range. This automatic assignment reflects the correlations between the variables. The correlation between CIELAB and CIEDE2000 is higher than the correlation between the triplet matching kernel and color names, and the assigned shapes reflect this relationship perceptually. For example, the perceptual distance between upward- and downward-pointing triangles is smaller than the perceptual distance between circle and square.

In a second example, we use visual embedding to encode community clusters in a character co-occurrence graph derived from Victor Hugo's novel Les Misérables. Cluster memberships were computed using a standard modularity-based community-detection algorithm (see [15]). For the data space distances, we count all inter-cluster edges and then normalize by the theoretically maximal number of edges between groups. To provide more dynamic range, we re-scale these normalized values to the range [0.2,0.8]. Clusters that share no connecting edges are given a maximal distance of 1. We then perform separate visual embeddings using univariate color and shape kernels (both estimated using triplet matching). As shown in Figure 17, the assigned colors and shapes perceptually reflect the inter-cluster relations.

Fig. 17: Graph of character co-occurrences in Les Misèrables, with node colors and shapes automatically chosen via visual embedding to reflect connection strengths between community clusters.

8 Conclusion

We introduce *perceptual kernels*, perceptual distance matrices formed from aggregate similarity judgments. Through a set of crowdsourced experiments, we compare the use of different judgment tasks to estimate perceptual distances. We find that ordinal triplet matching — in which subjects are shown a triplet of stimuli and asked to choose which of two items is more similar to a designated reference — exhibit the least inter-subject variance, are less sensitive to subject count, and enable the most accurate prediction of bivariate kernels from univariate inputs. Pairwise Likert scale judgments also fare well, and involve faster and cheaper experiments than triplet comparisons. Spatial arrangement tasks, on the other hand, exhibit much higher variance and can produce results inconsistent with existing perceptual models. Based on these considerations, we recommend the use of triplet matching judgments unless prohibited by issues of time or cost. We demonstrate how perceptual kernels enable automated design by re-ordering palettes to enhance discriminability and using *visual embedding* [8] to assign visual stimuli to data points in a structure-preserving fashion.

Our results also have broader implications. Our analysis is relevant to the general problem of crowdsourcing similarity models [1, 20, 31, 44, 47], providing new evidence in support of triplet matching. The poor performance of spatial arrangement (SA) also has implications for existing visual analytics tools. Semantic interaction systems (e.g., ForceSPIRE [9]) use SA tasks to elicit domain expertise to drive modeling and layout. Our results suggest that this mode of interaction may engender significant variation among experts and provide insufficient expressiveness for high-dimensional relations. Such tools may benefit by incorporating alternative similarity judgment tasks.

With respect to future work, integrating perceptual kernels into visualization design tools is an important next step. Towards this aim, we have made our perceptual kernels and experiment source code publicly available at *anonymized URL*. While we focused on specific shape, color, and size palettes, we plan to incorporate additional stimuli in each of these perceptual channels. Moreover, we can collect data for other channels, such as opacity, orientation, and lightness. Future work should also explore techniques for scaling to larger palettes, such as partitioning and adaptive sampling [44].

Future research might also extend our approach to more situated contexts. In this work we used direct measurement types, but it is possible to derive perceptual similarities through indirect judgments, such as the time taken to complete low-level graph reading tasks. As visual variables don't live in isolation, how different contexts may bias judgment remains an important concern. Gathering similarity judgments within the presence of competing variables would be valuable for as-

sessing contextual effects. In the meantime, perceptual kernels provide a useful operational model for incorporating empirical perception data directly into visualization design tools.

REFERENCES

[1] S. Agarwal, J. Wills, L. Cayton, G. Lanckriet, D. Kriegman, and S. Belongie. Generalized non-metric multidimensional scaling. In *International Conference on Artificial Intelligence and Statistics*, 2007.

[2] F. Attneave. Dimensions of similarity. *The American Journal of Psychology*, 63:511–556, 1950.

[3] J. Bertin. *Semiology of graphics*. University of Wisconsin Press, 1983.

[4] S. M. Casner. Task-analytic approach to the automated design of graphic presentations. *ACM Trans. Graph.*, 10(2):111–151, Apr. 1991.

[5] W. S. Cleveland. *Visualizing Data*. Hobart Press, 1993.

[6] W. S. Cleveland and R. McGill. Graphical perception: Theory, experimentation, and application to the development of graphical methods. *Journal of the American Statistical Association*, 79(387):531–554, 1984.

[7] Ç. Demiralp and D. H. Laidlaw. Similarity coloring of DTI fiber tracts. In *Proc. Med. Image. Comput. Comput. Assist. Interv. (MICCAI) Workshop on DMFC*, 2009.

[8] Ç. Demiralp, C. E. Scheiddegger, G. L. Kindlmann, D. H. Laidlaw, and J. Heer. Visual embedding: A model for visualization. *IEEE Computer Graphics & Applications*, 2014.

[9] A. Endert, P. Fiaux, and C. North. Semantic interaction for visual text analytics. In *ACM Human Factors in Computing Systems (CHI)*, pages 473–482, 2012.

[10] G. Fechner. *Elements of Psychophysics*. Holt, Rinehart and Winston, 1966.

[11] W. R. Garner and G. L. Felfoldy. Integrality of stimulus dimensions in various types of information processing. *Cognitive Psychology*, pages 225–241, 1970.

[12] R. Goldstone. An efficient method for obtaining similarity data. *Behavior Research Methods Instruments & Computers*, 26(4):381–386, 1994.

[13] J. Heer and M. Agrawala. Multi-scale banking to 45 degrees. *IEEE Trans. Visualization & Comp. Graphics*, 12:701–708, 2006.

[14] J. Heer and M. Bostock. Crowdsourcing graphical perception: Using mechanical turk to assess visualization design. In *ACM Human Factors in Computing Systems (CHI)*, 2010.

[15] J. Heer, M. Bostock, and V. Ogievetsky. A tour through the visualization zoo. *Communications of the ACM*, 53(6):59–67, June 2010.

[16] J. Heer, N. Kong, and M. Agrawala. Sizing the horizon: The effects of chart size and layering on the graphical perception of time series visualizations. In *ACM Human Factors in Computing Systems (CHI)*, 2009.

[17] J. Heer and G. Robertson. Animated transitions in statistical data graphics. *IEEE Trans. Visualization & Comp. Graphics*, 13:1240–1247, 2007.

[18] J. Heer and M. Stone. Color naming models for color selection, image editing and palette design. In *ACM Human Factors in Computing Systems (CHI)*, 2012.

[19] D. G. Kendall. A survey of the statistical theory of shape. *Statistical Science*, 4(2):pp. 87–99, 1989.

[20] M. Kendall. *Rank correlation methods*. Theory and applications of rank order-statistics. Hafner Pub. Co., 1962.

[21] N. Kong, J. Heer, and M. Agrawala. Perceptual guidelines for creating rectangular treemaps. *IEEE Trans. Visualization & Comp. Graphics*, 16(6):990–998, 2010.

[22] J. Kruskal. Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika*, 29(1):1–27, 1964.

[23] H. Lam, T. Munzner, and R. Kincaid. Overview use in multiple visual information resolution interfaces. *IEEE Trans. Visualization & Comp. Graphics*, 13(6):1278–1285, 2007.

[24] S. Lewandowsky and I. Spence. Discriminating strata in scatterplots. *Journal of American Statistical Association*, 84(407):682–688, 1989.

[25] A. MacEachren. *How Maps Work: Representation, Visualization, and Design*. Guilford Press, 1995.

[26] J. Mackinlay, P. Hanrahan, and C. Stolte. Show me: Automatic presentation for visual analysis. *IEEE Trans. Visualization & Comp. Graphics*, 13(6):1137–1144, 2007.

[27] J. D. Mackinlay. Automating the design of graphical presentations of relational information. *ACM Trans. Graph.*, 5(2):110–141, 1986.

[28] R. Munroe. Color survey results. http://blog.xkcd.com/2010/05/03/color-survey-results/, May 2010.

[29] G. Robertson, R. Fernandez, D. Fisher, B. Lee, and J. Stasko. Effectiveness of animation in trend visualization. *IEEE Trans. Visualization & Comp. Graphics*, 14(6):1325–1332, 2008.

[30] S. F. Roth, J. Kolojejchick, J. Mattis, and J. Goldstein. Interactive graphic design using automatic presentation knowledge. In *ACM Human Factors in Computing Systems (CHI)*, pages 112–117, 1994.

[31] M. Schultz and T. Joachims. Learning a distance metric from relative comparisons. In *Advances in Neural Information Processing Systems (NIPS)*. MIT Press, 2003.

[32] R. Shepard. The analysis of proximities: Multidimensional scaling with an unknown distance function. I. *Psychometrika*, 27(2):125–140, 1962.

[33] R. N. Shepard. Attention and the metric structure of the stimulus space. *Journal of Mathematical Psychology*, pages 54–87, 1964.

[34] R. N. Shepard. Toward a Universal Law of Generalization for Psychological Science. *Science*, 237(4820):1317–1323, 1987.

[35] R. N. Shepard. Integrality versus separability of stimulus dimensions: From an early convergence of evidence to a proposed theoretical basis. In *The perception of structure: Essays in honor of Wendell R. Garner*. APA, 1991.

[36] B. Shortridge. Stimulus processing models from psychology: can we use them in cartography? *The American Cartographer*, 9:155–167, 1982.

[37] D. Simkin and R. Hastie. An information-processing analysis of graph perception. *Journal of American Statistical Association*, 82(398):454–465, 1987.

[38] I. Spence and S. Lewandowsky. Displaying proportions and percentages. *Applied Cognitive Psychology*, 5:61–77, 1991.

[39] S. S. Stevens. On the psychophysical law. *Psychological Review*, 64:153–181, 1957.

[40] M. Stone. Color in information display. http://www.stonesc.com/VisCourses.htm.

[41] M. Stone and L. Bartram. Alpha, contrast and the perception of visual metadata. In *Color Imaging Conference*, 2009.

[42] J. Talbot, J. Gerth, and P. Hanrahan. Arc length-based aspect ratio selection. *IEEE Trans. Visualization & Comp. Graphics*, 2011.

[43] J. Talbot, S. Lin, and P. Hanrahan. An extension of Wilkinson's algorithm for positioning tick labels on axes. *IEEE Trans. Visualization & Comp. Graphics*, 2010.

[44] O. Tamuz, C. Liu, O. Shamir, A. Kalai, and S. J. Belongie. Adaptively learning the crowd kernel. In L. Getoor and T. Scheffer, editors, *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, pages 673–680. ACM, 2011.

[45] W. Torgerson. Multidimensional scaling: I. theory and method. *Psychometrika*, 17(4):401–419, 1952.

[46] L. Tremmel. The visual separability of plotting symbols in scatterplots. *Journal of Computational and Graphical Statistics*, 4(2):101–112, 1995.

[47] L. van der Maaten and K. Weinberger. Stochastic triplet embedding. In *Machine Learning for Signal Processing (MLSP), 2012 IEEE International Workshop on*, pages 1–6, 2012.