https://doi.org/10.2352/ISSN.2169-2629.2018.26.365

Single Anchor Sorting of Visual Appearance as an Oriented Graph

N. Moroney, I. Tastl and M. Gottwals; HP Labs, HP Inc.; Palo Alto, CA, USA
M. Ludwig and G. Meyer; University of Minnesota; MN, USA

Abstract

Given a single reference stimulus, test stimuli can be sorted with respect to perceptual similarity to this anchor stimulus. Aggregated ranks can then be computed from multiple sort sequences. This ordinal scaling provides an estimate of perceptible differences and can be used to develop and test predictive models. In this paper we propose the use of graph-based methods visualizing experimental data and computing aggregated ranks. Specifically, perceptual similarity is expressed as a sort sequence graph in which nodes are stimuli and weighted edges are the frequency of the corresponding ranks. This graph is also oriented in that it has a start, the reference stimuli, and an end, the least similar stimuli. The Schulze method or the 'strongest path' computation is used for rank aggregation. This analysis is explored in the context of two appearance experiments: the first using solid colors and the second using renderings of 3D printed stimuli varying in multiple appearance attributes. For the second experiment with the renderings of 3D printed stimuli we then use Kendall T_b values to assess a simple model based on mean CIELAB color differences. We find that the underlying sorting task is efficient and intuitive. Furthermore, the graph-based formulation of perceptual similarity allows the application of network analysis and graph theory to the study of visual appearance. New analyses are also possible, such as outlier detection using the sort sequences that are the inverse of the Schulze solution or approximately the 'wrongest path'.

Introduction

Ranking of stimuli or the sorting of visual appearance is a familiar technique for psychometric experimentation. For example, in Engeldrum [1] the derivation of a rank-stimulus matrix from ranking data is described, as is the computation of an average rank. It is also noted that this is typically applicable to a single "-ness" or attribute such as glossiness or lightness. Another familiar color science example, is the Farnsworth-Munsell 100 Hue Test [2]. This test in which color samples are sorted, twenty-five at a time, in order of hue given two end points uses disagreements in the observer sorting as an indication of anomalous color vision. Additional published research results that make use of sorting includes Bimler[3], Rogowitz [4], Moroney [5] and Clarke [6]. Our contributions include, focusing on use of a single reference anchor, transformation of raw observer rank data into a sort sequence graph and the use of the Schulze method to perform rank aggregation of the perceptual ranking. This introduction section provides a single simplified description of each of these and is then followed with a more detailed description of the two visual experiments.

For the visual sorting of stimuli, consider six randomized stimuli arranged in an arc. A single reference stimulus is presented in the center and the observer's task is to sort the stimuli from more similar on the left to less similar on the right of the arc. This was implemented using a drag-and-drop user interface design and the result is a quick and intuitive method of manipulating stimuli similarity. For both experiments, a test task of sorting shapes was used. An initial, randomized screen shot of this task is shown in the top of Figure 1. In this case, the reference is a circle and a range of shapes is provided as test stimuli. One observer's sorting of these shapes is shown at the bottom of Figure 1. In this case, the observer selected a thick ellipse as being most similar to the reference by placing it furthest to the left. By convention this will be the rank 1 position. Likewise, the observer selected the triangle as being least like the circle by placing it furthest to the right. This location will also be referred to as the rank 6 position.

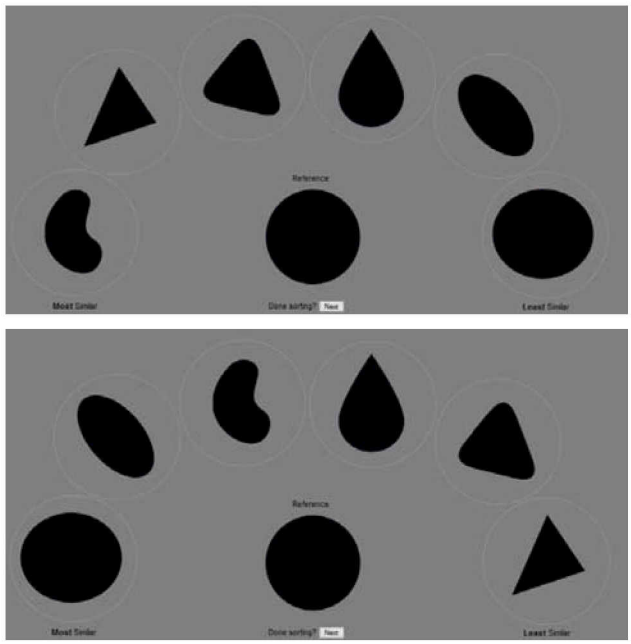

Figure 1. Unsorted shapes(top) and shapes sorted by one observer by similarity relative to central stimulus (bottom). The observer was instructed to sort from more to less similar using a left-to-right orientation.

Given multiple stimuli rankings, either from multiple observers or from observer repetitions, the next step is the transformation of raw ranks to a **sort-sequence graph**. This is a graph consisting of stimuli at the nodes. The weighted edges are the frequency of nodes having a corresponding rank sequences for all the observations. As a simple example assume three stimuli A, B and C with two sort sequences of A,B,C and B,A,C. There are two edges for AB, one for BC and one for AC. There would also be an edge between the reference and A and the reference and B. Finally, there will be two edges between C and the least similar end of graph node. These are directed edges but for simplicity will be shown as undirected edges.

Given a weighted graph, force-based layout algorithms, such as the Fruchterman-Reingold [7] algorithm are useful for generating an initial layout of the sort-sequence graph. As the graph is constructed using a progression of ranks we also manually orient the graphs such that rank 1 or most similar stimuli is on the left while rank 6 or least similar stimuli are on the right. We include the reference stimuli to the left of rank 1 and an empty end of sequence node to the right of rank 6. While the left-right association of number lines to numerosity is subject to ongoing research [8], we use left-more similar and right-less similar for the sort sequence graphs to be consistent with the orientation used in the visual sorting task. Given a sort-sequence graph, it is then possible to make comparisons to the literature of graph theory and network analysis.[9] For example, if the shape stimuli shown in Figure 1 were indistinguishable then the result would be a random graph. In contrast, if the shape stimuli had a single, unambiguous rank sequence then the result would be a linear arrangement of edges with only the nearest neighbor connections. Abstract examples of both the random and linear-arrangement graphs are shown in Figure 2.

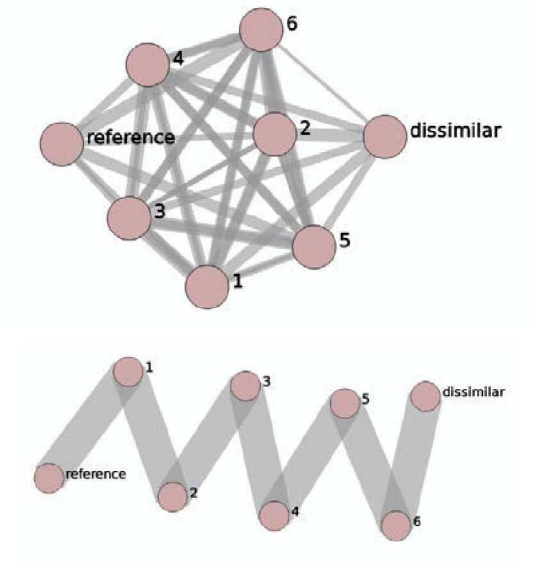

Figure 2. A simulated random sort-sequence graph (top) and a simulated perfect or universal agreement sort-sequence graph (bottom).

For the test case of black-and-white shapes shown in Figure 1, the sort sequence with outliers removed is shown in Figure 3. The specifics of the outlier removal will be discussed in a later section. The results shown in Figure 3 differ from either of the graphs shown in Figure 2. The shape stimuli neither follow a random nor a linear arrangement. While there is universal agreement from the observers that the thicker ellipse, followed by the thinner ellipse are most similar to the anchor circle shape, there is less agreement after that. Note that this disagreement is likely not due to approaching a perceptibility threshold but is instead a reflection of multiple cognitive sorting criteria [10] by individual observers. This disagreement is also not random. For example, none of the observers provided a sorting in which the thinner ellipse and the triangle with rounded corners were neighbors.

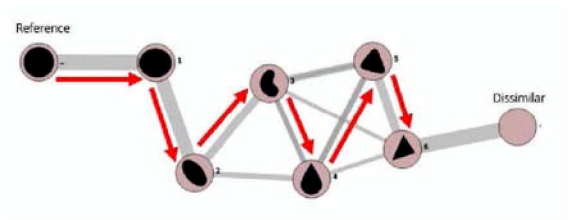

Figure 3. The sort sequence graph for 150 observers for the shape stimuli shown in Figure 1. This graph can be visually compared to the simulated random and perfect sort-sequence graphs shown in Figure 2.

Note that the thickness of the edges in the sort sequence graph shown in Figure 3, corresponds to the proportion of observers ranking the stimuli as neighbors. It is shown as an undirected graph for simplicity, but the underlying data is directed. Given a directed graph of ranks or preferences it is possible to compute aggregated ranks for the stimuli using the Schulze method. This voting technique essentially computes a single ranking based on the 'strongest path' through the stimulus sort sequence graph. This is shown as a series of red arrows in Figure 4 and was computed using the method described in reference [6]. It is also possible to use the inverse of this concept, that is the sort sequence that is the opposite of the strongest path, or the 'wrongest path' as a means of detecting outliers. This data can indicate lack of instructional clarity or adversarial participants. In either case this is one of the outlier detection techniques used in this paper.

Experimental Interface

For the collection of stimulus rank sorting data, we used a simple web-based program. An HTML5 drag-and-drop interface was implemented and the test stimuli, instructions and data collection were performed in the browser. The stimuli were scaled in size to fill an average desktop display or larger than a 10-degree angular subtense. The experiment started with an instruction screen and an example gif showing stimuli being sorted by size. At the beginning the observer had unlimited time to sort the stimuli. After sorting a subset or in the laboratory study, all stimuli, the observers were thanked for their participation and a link for more background information was provided. For the solid color sorting experiment in the laboratory, an sRGB calibrated display was used in a dark room. For the rendered 3D printed experiment, data was collected using the internet. For all participants and all sorting trials, the median time to complete the task was 24 seconds and consisted of a median of 5 drags to sort the stimuli. All data was collected anonymously and voluntarily. No cookies were used, nor was any other personal data collected. Each block of experimental data was assigned a random unique user identifier at the time of participation but is not re-used or otherwise associated with participants.

Solid Color Sorting Experiment

For the solid color sorting experiment, the stimuli consisted of a circular region with a single uniform color. The only attribute that varied was color. In addition to the shape sorting stimulus set shown in a Figure 1, there was a second test case used at the beginning of each experiment. This test case was a sampling of colors along the deuteranopic confusion line. [12] This sampling was computed assuming an sRGB display and using the Smith-Pokorny copunctal

points. [13] Color vision deficiencies are a challenge for collecting visual appearance data using the internet. In this case there is an underlying model representing a sorting of these stimuli by observers that do not have a color deficiency. This provides an additional means of weighting the data or performing outlier detection. The encouraging results are that a large majority of participants provided a perfect sorting of the deuteranopic sampling. Applying additional outlier detection, or removal of rankings that mixed categories or intermixing reds with greens results in the sort sequence graph shown in Figure 4. Given the universal agreement in the sort sequence, rank aggregation is trivial and there is perfect agreement with the sort sequence predicted by CIELAB color differences.

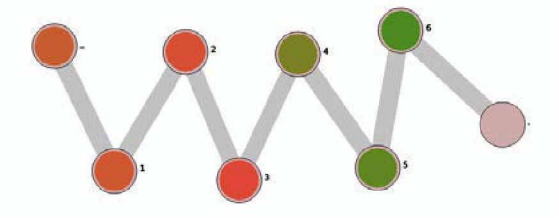

Figure 4. The sort sequence for the deuteranopic confusion sampling, used as first of two test screens for the web-based sorting experiment. This result shows a sequence that, with outliers removed, shows perfect agreement with the color differences.

In addition to this test case of deuteranopic confusion sampling, additional laboratory data was collected for a set of solid color stimuli. Five reference colors were selected based on a nominal data set: gray, brown, olive, teal and pink. Given these centroids, several fixed monotonic step sequences were computed using ΔE*ab. Stimuli with steps sizes of 3, 2, 1 and 0.5 ΔE*ab were generated. MATLAB's symbolic toolbox was used to achieve step ramps of desired size and orientation. Ten different univariate and bivariate changes were computed for the step ramps with constant steps in either L*, a*, b* or equal changes in a* and b*. In all cases samples with gamut clippings at the sRGB boundary were avoided and the corresponding sRGB values were used. An a* versus b* plot of the centroids and a sub-set of the stimuli is shown in Figure 5. Four observers completed the sort experiment for 3, 2, and 1 ΔE*ab and one observer also did a 0.5 ΔE^*_{ab} sort experiment. An HP DreamColor Z32x display was used in sRGB mode in a dark viewing environment.

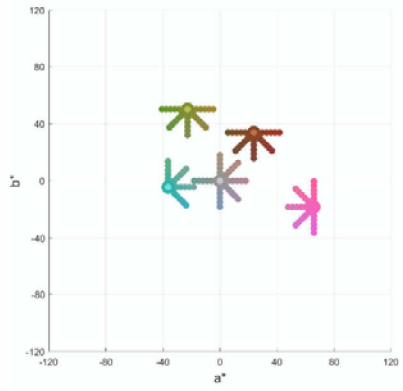

*Figure 5. Color centroids and example constant step-size stimuli generated for the constant 3 ΔE*ab trial of the solid color sorting experiment shown in a* versus b* plot in the CIELAB color space.*

The aggregated sort sequence graphs for the 1 ΔE^*_{ab} samplings for the gray, cyan and magenta colors are shown in Figure 6. These results can be compared with the previous sort sequence graphs. Note that all the nodes are shown using the reference color since these graphs are aggregated across all the sampling paths shown in Figure 5 and not enough data was collected to provide sort-sequence graphs for each individual ramp. Qualitatively, they demonstrate that the grays or achromatic samples were almost perfectly sorted with respect to the original CIELAB sampling, with only nearest neighbor disagreements in rankings of the color differences. These results are in contrast to the pink sort sequence graph which is approaching the random graph shown Figure 2. The results for the cyan colors are intermediate to the grays and pinks in that there is an evident agreement with the underlying rankings but their differences in ranks are larger than just nearest neighbors. Additional data collection is ongoing and relevant approaches to analysis will be considered in the next section.

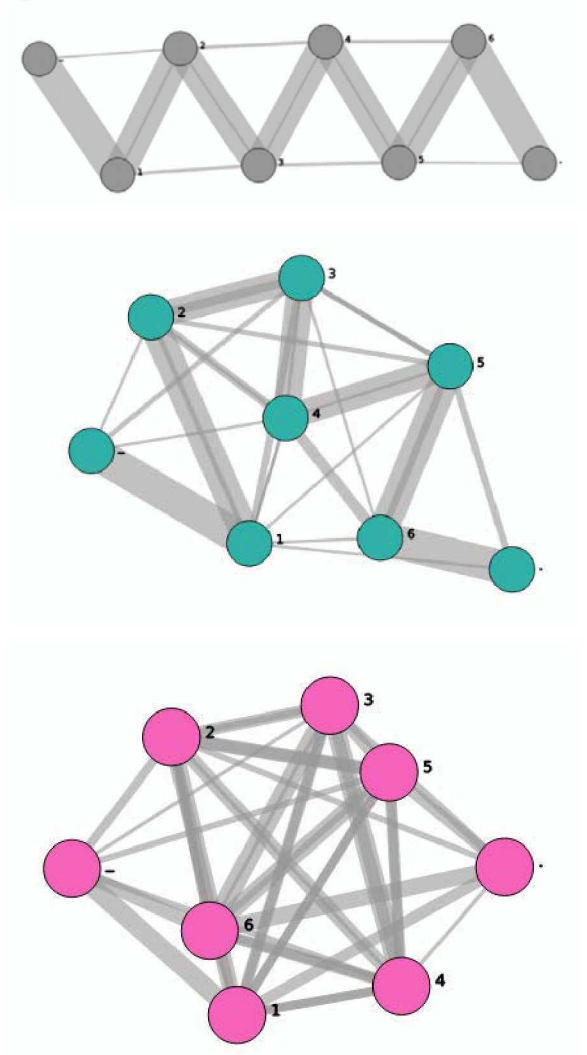

*Figure 6. The sort sequence for the gray(top), cyan (middle) and pink (bottom) solid colors sampled at 1 DE*ab.*

Rendered 3D Print Sorting Experiment

The second sorting experiment performed made use of renderings of the 3D printed appearance of a more complex test stimulus [14], referred to here as the blob. This experiment builds on previous experimental results and is part of a broader effort to derive a more comprehensive model for the appearance of 3D printed objects. As such this paper focuses on the use of web-based ranking to collect ground truth data.

Thirty unique trials were prepared where the stimuli images were physically-based renderings of a 3D blob featuring various complex appearances. Each task consisted of a reference appearance and six candidates to be reordered by a subject. The appearances were produced from 3D printing colored tiles (cyan, green, and purple) printed at several orientations (tiles that were placed at 0, 10, and 45 degrees to the z-axis of the 3D printer) and who's material appearance properties were captured using an X-Rite TAC7 scanning device.

Additional appearances were scanned after post-processing the printed tiles to create smoother and glossier variants that had similar spatial patterns. The candidates in a sorting task represent modifications of the reference appearance. The modifications group the 30 tasks into sets comparing:

1. *Roughness against diffuse contrast*: the reference was virtually modified to produce three rougher versions at 0.1, 2.0, and 4.0 scaled height profiles; and three contrast versions at 0.1, 2.0, and 3.0 scaled Weber contrast of the diffuse color.
2. *Contrast against printed angle*: the Weber contrast was scaled by 0.1, 2.0, and 3.0 for the two printing orientations not matching the reference (e.g. if the reference appearance was printed at 10 degrees, three candidates come from 0 degrees and the remaining three are from 45 degrees).
3. *Roughness against post-processing technique*: three height-scaled versions were generated from the respective original tiles and from the matching post-processed tiles.
4. *Contrast against post-processing technique*: three Weber contrast variants were generated from the respective original tiles and from the matching post-processed tiles.

The first and second group each had nine reference appearances, covering the combination of three colors and three printed orientations of the original printed tiles. In those groups, the post-processed modifications were not evaluated. The third and fourth group each had six references: cyan at 0 degrees, green at 10 degrees, and purple at 45 degrees, using both original and post-processed variants. Figure 7 shows the reference objects or renderings used for this experiment.

The structure of these thirty tasks force the subjects to evaluate and rank the six candidates that represent transformations along two complex appearance dimensions. The specific transformations are not known to the subjects and are only perceivable based on how the changes affect the physically-based renderings. It is not uncommon for changes in roughness, which result in increasing shadows and breaking apart of smooth specular highlights, to induce visual changes that are similar to those produced by increasing the diffuse contrast of the material.

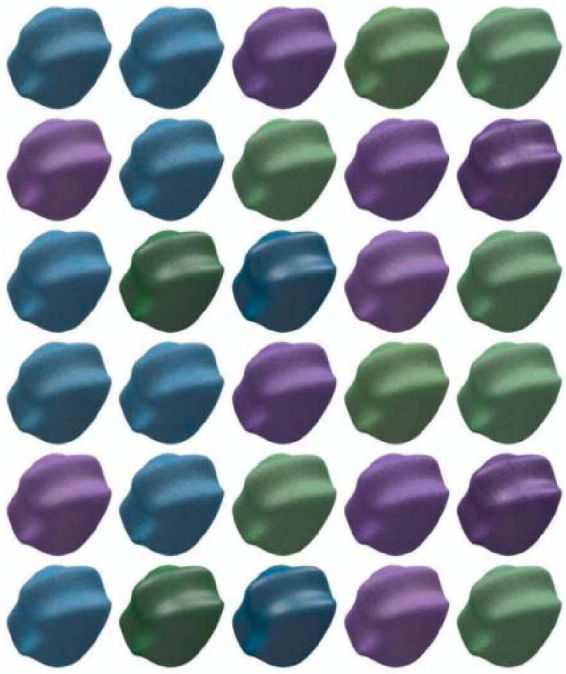

Figure 7. Reference objects or renderings used for the complex appearance sorting experiment. Color appearances were simulated based on 3D printed and measured reference materials.

The complex appearance sorting experiment was conducted online and data was collected from over 150 participants. Each volunteer started by sorting the shape and deuteranopic color sorting tests, described earlier. Subsequently, they then sorted 5 randomly selected set of blobs. This resulted in an average of 20 volunteers sorting a specific set of blobs.

The results for the blob sorting experiment are shown in two forms. One is using an example sort sequence graph for one of the trials. Second is a summary plot for all trials which shows the resulting performance of using the mean ΔE *ab for assessing the differences in appearance between the blobs. Figure 8 shows the aggregated sort sequence for a cyan reference.

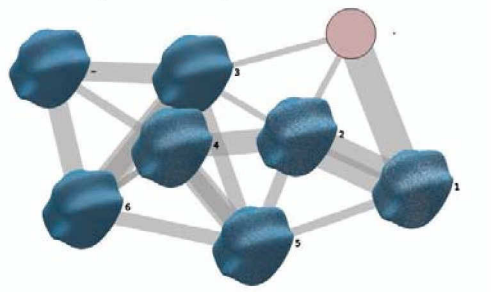

Figure 8. One of the 30 sort sequences for the blobs or complex appearance sortings. The reference is shown on the '-' node to the upper left while the ending node for the end of sequence or neighbor to the least similar stimuli is shown as an '.' node with no corresponding image.

The aggregated sort sequence graph shown in Figure 8 can be analyzed using the Schulze method for finding the strongest path. This then yields an aggregated rank. This analysis was repeated for all 30 blob sets and the resulting aggregated ranks can be used as a ground truth. As a simple test case, consider the average ΔE *ab for all the pixels in the stimulus blob relative to the reference bob. This set of color differences can then be converted to a ranking and compared to the Schulze aggregated ranks. To compare the rank correlations, the Kendall T_a values can be computed [15] for each of the blob sets. To aid the visualization, the blob sets can be sorted by the Kendall T_a values such that the worst correlation is to the left and the best is to the right. This is shown in Figure 9 and the plotted data points are supplemented with two light gray horizontal lines at the plus/minus 1 standard deviation for the random sorter. This figure then shows how well a computational model, such as mean ΔE *ab, can work relative to an abstract model, such as random sorting. It also demonstrates the complete process of collecting complex appearance sorting data, performing rank aggregation using the Schulze method and finally performing model assessment using Kendall T_a rank correlations.

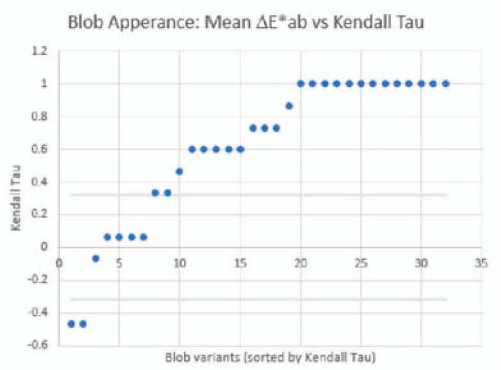

Figure 9. Summary plot of blob appearance sorting experiment. The x-axis shows the blob variants or experimental trial as sorted by the Kendall T_a. The y-axis is the as computed Kendall T_a between the similarity sorted stimuli and the rank as predicted by the CIELAB color difference.

The results shown in Figure 9 have a median correlation of 0.73 and the sorted trials allow a visual inspection of which trials or stimulus sets the color difference performed well, or as importantly poorly. The use of the blob CIELAB color difference is used not as a proposal that it be used for complex appearance analysis, in fact the authors have considered more comprehensive models for this and related data. [16,17] Instead it is used as a familiar metric that provides a complete instance of similarity sorted stimuli to model evaluation. To this end the sorting data and blob stimuli are available at the following source code repository. [18]

Discussion

This paper has described and used rank sequence graphs to represent and analyze the results of single anchor sorting experiments. Relatively established metrics, such as the widely known ΔE *ab, were used on purpose to elucidate the method and there is ongoing work on using the data to refine, extend or formulate models. For example, for the solid color sorting experiment, more advanced color difference metrics could be compared to the results of ΔE *ab using processing comparable to that described for the blob appearance modeling with ΔE *ab in Figure 5. Likewise, use of more extensive complex appearance models can be used and compared.

The previous sections have also mentioned outlier detection as another benefit of the rank sequence graphs and strongest graph analysis. Specifically given a collection of ranks, such as for the deuteranopic confusion colors, it is possible to use the strongest path the find it's opposite. That is, given the strongest path find the path or paths with the least rank correlation as possible 'wrongest' paths. This type of analysis was applied in Figures 8 and 9 and while a simple process, depends on the rank sequence graph. Use of a rank-stimulus matrix and average ranks does not yield a comparable form of outlier detection.

Conclusions

This paper represents and quantifies single anchor visual sorting of appearances using rank sequence graphs. Rank aggregation was computed using the Schulze method and experimental results were presented for solid color sampling and complex visual appearance. Solid color and complex object appearance stimuli were sorted using two web-based experiments. The results show that graph-based analysis can be used for both visualization of the raw observer data and for model evaluation. In the case of complex object appearance, the Kendall T_a rank correlations between observer Schulze aggregated ranks is used to assess the predictive power of CIELAB color differences.

References

[1] Engeldrum, Peter G. Psychometric scaling: a toolkit for imaging systems development. Imcotek, 2000.

[2] Farnsworth, Dean. "The Farnsworth-Munsell 100-hue and dichotomous tests for color vision." JOSA 33, no. 10 (1943): 568-578.

[3] Bimler, David and John Kirkland, "Perceptual Modelling of Product Similarities Using Sorting Data". Marketing Bulletin, 9, pp. 16-27 (1998).

[4] Rogowitz, Bernice E., et al. "Perceptual image similarity experiments." Human Vision and Electronic Imaging. Vol. 3. No. 3299. 1998. In this experiment a full 2D sort of 97 printed images.

[5] Moroney Nathan, Ingeborg Tastl, and Melanie Gottwals. "A Similarity Measure for Large Color Differences." Color and Imaging Conference. Vol. 2014. No. 2014. Society for Imaging Science and Technology, 2014.

[6] Clarke, Alasdair DF, Xinghui Dong, and Mike J. Chantler. "Does free-sorting provide a good estimate of visual similarity." Predicting Perceptions (2012): 17-20. Free-sorting of 334 texture images by 30 participants.

[7] Fruchterman, Thomas MJ, and Edward M. Reingold. "Graph drawing by force-directed placement." *Software: Practice and experience* 21.11 (1991): 1129-1164.

[8] de Hevia, Maria Dolores, et al. "Human infants' preference for left-to-right oriented increasing numerical sequences." PloS one 9.5 (2014)

[9] Newman, M.E.J, Networks: An Introduction, Oxford University Press, Oxford, UK, 2010.

[10] Regehr, Glenn, and Lee R. Brooks. "Category organization in free classification: The organizing effect of an array of stimuli." Journal of

Experimental Psychology: Learning, Memory, and Cognition 21, no. 2 (1995): 347.

[11] Schulze, Markus. "A new monotonic, clone-independent, reversal symmetric, and condorcet-consistent single-winner election method." Social Choice and Welfare 36, no. 2 (2011): 267-303.

[12] Meyer, Gary W., and Donald P. Greenberg. "Color-defective vision and computer graphics displays." IEEE Computer Graphics and Applications 8.5 (1988): 28-40.

[13] Shevell, Steven K., ed. The science of color. P. 140 Elsevier, 2003.

[14] Fleming, Roland W., Dror Ron O.,, and Adelson, Edward H.. 2003. Real-world illumination and the perception of surface reflectance properties. Journal of Vision 3, 5 (June 2003), 347–368.

[15] Cliff, Norman. Ordinal methods for behavioral data analysis. Psychology Press, 2014.

[16] Ludwig Michael, Moroney Nathan, Tastl Ingeborg, Gottwals Melanie, and Meyer Gary. "Perceptual Appearance Uniformity in 3D Printing". In EI 1-12, https://doi.org/10.2352/ISSN.2470-1173.2018.8.MAAP-209.

[17] Ludwig, Michael, Gary Meyer, Ingeborg Tastl, Nathan Moroney, and Melanie Gottwals. "An appearance uniformity metric for 3D printing." In Proceedings of the 15th ACM Symposium on Applied Perception, p. 14. ACM, 2018.

[18] https://github.com/NMoroney/SimilaritySorting

Author Biographies

Nathan, Ingeborg and Melanie are with HP Inc. and are researchers in the Print Adjacencies and 3D Lab at HP Labs in Palo Alto, CA. Their current efforts relate to color appearance for 3D printing. Michael completed his PhD at the University of Minnesota where his advisor, Gary, is an Associate Professor in the Department of Computer Science and Engineering. Michael is currently with Google in North Carolina.