Erschienen in: ACM Transactions on Applied Perception (TAP) ; 16 (2019), 1. - 5
https://dx.doi.org/10.1145/3301414

Quantifying Visual Abstraction Quality for Computer-Generated Illustrations

MARC SPICKER, FRANZ GÖTZ-HAHN, THOMAS LINDEMEIER, DIETMAR SAUPE, and OLIVER DEUSSEN, University of Konstanz, Germany

We investigate how the perceived abstraction quality of computer-generated illustrations is related to the number of primitives (points and small lines) used to create them. Since it is difficult to find objective functions that quantify the visual quality of such illustrations, we propose an approach to derive perceptual models from a user study. By gathering comparative data in a crowdsourcing user study and employing a paired comparison model, we can reconstruct absolute quality values. Based on an exemplary study for stippling, we show that it is possible to model the perceived quality of stippled representations based on the properties of an input image. The generalizability of our approach is demonstrated by comparing models for different stippling methods. By showing that our proposed approach also works for small lines, we demonstrate its applicability toward quantifying different representational drawing elements. Our results can be related to Weber–Fechner's law from psychophysics and indicate a logarithmic relationship between number of rendering primitives in an illustration and the perceived abstraction quality thereof.

CCS Concepts: • **Computing methodologies** → **Non-photorealistic rendering**; **Perception**;

Additional Key Words and Phrases: Visual abstraction, user study, perception, stippling, non-photorealistic rendering

1 INTRODUCTION

Inspired by artistic and expressive styles, the field of Non-Photorealistic Rendering (NPR) focuses on the automatic creation of abstract renditions with drawing primitives, such as dots, lines, and textured strokes. A vast amount of research has been dedicated to the automatic creation of such illustrations. An example for this is stippling, a powerful illustration technique using only dots, which can frequently be found in areas like archeology and biology (Hodges 2003). Using interactive systems can greatly reduce the time required by an artist to create such illustrations (Deussen et al. 2000). Techniques that are capable of running at interactive frame rates have also been around for some time (Pastor et al. 2003). Many of these methods aim to optimize the blue

The authors thank the German Research Foundation (DFG) for financial support within projects A04 and A05 of SFB/Transregio 161.
Authors' addresses: M. Spicker, F. Götz-Hahn, T. Lindemeier, D. Saupe, and O. Deussen, University of Konstanz, Department of Computer and Information Science, Universitätsstraße 10, Konstanz, 78457, Germany; emails: marc.spicker@googlemail.com, {hahn.franz, thomas.lindemeier}@gmail.com, {dietmar.saupe, oliver.deussen}@uni-konstanz.de.

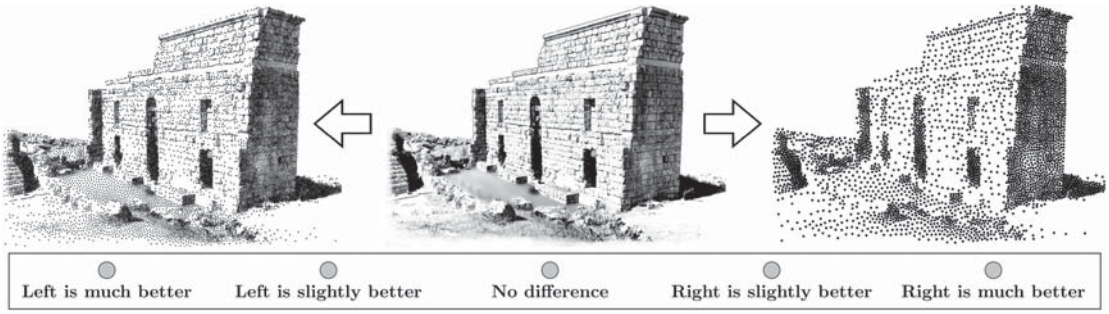

Fig. 1. We reconstruct absolute perceived quality scores for abstract visual representations such as stippling from relative comparisons and show that it is possible to predict this quantity solely based on properties of an input image.

noise properties of the point sets, which is commonly argued to be a quality metric for stippling and pointillism. Example-based approaches, e.g., Kim et al. (2009), combine human input in form of drawn tonal maps with fast automatic point placements. So far, computer-generated stipple illustrations have only been evaluated with regards to how good they resemble artworks created by hand (Isenberg et al. 2006; Maciejewski et al. 2007, 2008).

In contrast to such studies, this article focuses on the perceptual evaluation of illustrations with regards to the quality of abstraction. For this, we limit ourselves to techniques using countable drawing primitives, investigating how the number, size, and distribution of elements is related to the perceived quality of an illustration. An answer to this question is crucial, because most algorithms require the user to manually select the number of primitives, which is usually a trade-off between representativeness, computation time, and the problem that with too many elements the illustration loses its specific look. Our goal is to help users make an educated choice about this number and allow them to estimate the abstraction quality based on our findings.

Since there is no objective function to judge human perception of such drawings, we propose an approach to deduce perceptual models from a user study. Users are presented with the input image along with two illustrations created with a different number of primitives, with the task to judge which one is perceived as the better abstract representation. From many such comparisons it is possible to infer an absolute abstraction quality scale using Thurstone's model (Woods et al. 2010). While we present a general approach for representative drawing elements, we illustrate its use for stippling. We use stippling because of its small and therefore manageable parameter space (position and size). The main contributions in this article are the following:

- We introduce a tonal percentage measure for comparing drawings created with distinct rendering elements from different inputs.
- We propose an approach to derive perceptual models from a user study by gathering comparative data and employing a paired comparison model.
- We perform studies that assess the subjectively perceived quality of stippled abstractions, showing that the tonal percentage is related to the perceived quality by a log-like behavior and give guidance for deciding the number of primitives to be used.
- We test the generalizability and applicability of our approach by comparing models for different point distribution strategies for stippling and for one line-based abstraction technique.

This article represents an extended journal version of the NPAR 2017 paper by Spicker et al. (2017). Compared to the original paper, the additional contributions are the following: First, we now include a more advanced point distribution method for variable point sizes and compare both stippling models based on direct comparison studies. Second, we also apply our approach to a different rendering primitive with lines for hatching to test the

generalizability of our approach. This work is structured the same as the original paper: After reviewing related works, we introduce the input normalization approach and describe our illustration method. Next, we outline the theory that describes how we reconstruct absolute scale values from paired comparison data and discuss our user studies, from design and quality control to analysis aspects, and describe our proposed models. Finally, we present the conclusions from our studies and motivate potential directions for future works.

2 RELATED WORK

The automatic creation of abstract illustrations using distinct drawing primitives can be reduced to the careful placement of elements. Here, the goal is to create a distribution that matches the appearance and shading of an input. In the monochromatic case, for example, a higher amount of black elements on a white background are required to simulate darker areas and vice versa. These elements are often distributed based on Lloyd's relaxation method (Lloyd 1982), which maximizes distances between adjacent elements. Hiller et al. (2003) extend Lloyd's method to distribute arbitrary objects on a plane. While the original algorithm solely moves the objects, they also modify their orientation.

Stippling Techniques. The artistic process behind the creation of stipple drawings is analyzed by Deussen et al. (2000). They propose an interactive editor to create such drawings much faster, which is based on Lloyd's method. Secord (2002) builds upon this approach by extending it with weighted centroidal Voronoi diagrams, which adapt to the grayscale values of an input image. Kim et al. (2008) restrict the movement of cells in Lloyd's method at parallel offset lines, guiding stipples along image features. An energy-based term is introduced into the relaxation by Deussen (2009). Here, aesthetically pleasing point configurations can be related to specific amounts of energy. The blue-noise properties of point sets, a property commonly believed to describe the quality of point sets, is optimized by Balzer et al. (2009). They introduce the concept of capacity for each point, enforcing that each point obtains equal importance in the distribution. This allows to avoid visually unpleasing hexagonal substructures formed by the relaxation. An adaptive version of Lloyd's method inspired by the Linde–Buzo–Gray algorithm is proposed by Deussen et al. (2017). Voronoi cells are dynamically split and merged until a desired point density is reached. The algorithm is also capable to create results with variable point sizes.

Hand-drawn stippled illustrations are used by example-based methods to create a more faithful artistic reproduction: Kim et al. (2009) use texture synthesis and a similarity metric to reproduce perceptually similar stipple textures to those of an artist. Stippling is treated as a scale-dependent grayscale process by Martín et al. (2011). Scanned stipple examples steer their method depending on the spatial output size and resolution.

Stipple drawings have also been created using approaches that do not fit in one of the prior categories: Mould (2007) combines stippling and graph theory by transforming an input image into a weighted graph. Local image gradients are used as weights and the graph is traversed with Dijkstra's algorithm. Stipples are placed whenever the sum over traversed edges exceeds a given threshold.

Li and Mould (2011) provide a priority-based error diffusion method focusing on retaining image structure Their approach gives higher priority to extremal values, better retaining contrast compared to similar methods. Non-periodic point sets are generated by the technique of Kopf et al. (2006). These tileable sets enable them to create larger point sets from smaller building blocks. The speed gain allows viewers to resize stippled images with large numbers of points at interactive rates. de Goes et al. (2012) formulate the calculation of a Capacity-Constrained Voronoi Tesselation (CCVT) as an optimal transport problem. By enforcing capacity constraints exactly, they can create point sets with high-quality blue noise and spectral properties. For a more extensive discussion of stippling, we refer to the article of Martín et al. (2017) and the book of Deussen and Isenberg (2013).

Line Drawing Techniques. Winkenbach and Salesin (1994) propose an automated rendering system creating traditional pen-and-ink illustrations using stroke textures. Resolution-dependent strokes allow to stylize even complex architectural models. By analyzing a training patch of strokes, the approach of Jodoin et al. (2002) synthesizes sequences of strokes to resemble a target. It employs an extended texture synthesis methods that

works directly on parametrical curves instead of pixels. Singh and Schaefer (2010) construct a gradient field from the diffuse surface intensity of a model to guide a set of adaptively spaced lines representing the lighting situation under which the shape is viewed. An hierarchical proximity grid is used to improve the line quality and control their density.

An interactive real-time system giving users low-level control over stroke placement, while parameters such as tone, smudge, and details can be controlled on a higher-level, is presented by Durand et al. (2001). The system performs semi-automatic tonal modeling by applying a thresholding model of strokes. Based on the analysis of neighborhood relationships in an interactive system, Barla et al. (2006) synthesize different styles of stroke patterns. Starting from a user-specified reference pattern, a texture synthesis technique is applied that first determines meaningful pattern elements. A programmable approach for line drawings from 3D models similar to shaders in traditional rendering is proposed by Grabli et al. (2010). Users can freely define relations between style attributes and scene properties. A map describing support lines and their topological arrangement is created from the 3D model to ensure continuity of scene properties along edges.

Evaluations. The usage of artistic composition principles to improve the quality of abstract renderings is discussed by Rivotti et al. (2007). Two case studies investigate the influence of these principles, however they do not perform any quantitative evaluation. Cole et al. (2009) evaluate the effect of applying line drawing on the perception of three-dimensional objects, concluding that they can effectively depict shape. A study examining the understanding and assessment of hand-drawn pen-and-ink illustrations of objects compared to computer-generated illustrations is presented by Isenberg et al. (2006). They conclude that people perceived differences between the two, but both are still values as scientific illustrations. Maciejewski et al. (2007, 2008) explore the differences between human and computer-generated stipple illustrations using image-processing techniques. They show that the statistics of dot distributions, which influences aesthetics, varies between hand-drawn, computer-generated, and natural dot textures. Based on these statistics, man-made stipple drawings approximate natural textures more faithfully. To create a more faithful digital replication of the traditional artistic stippling process, Martín et al. (2015) focus on properties of artistic tools, such as pens and paper types. From the results of a user study, they provide a dataset for example-based stippling to create more faithful stipple drawings. For a more thorough overview over evaluations in the field of Non-Photorealistic Rendering, we refer to the works of Gatzidis et al. (2008) and Isenberg (2013). In information visualization, evaluations have shown that perceptual laws can be used to evaluate visualization designs, one of the examples being Weber's law (Harrison et al. 2014). A more general model has later been argued to be more accurate (Kay and Heer 2016).

The majority of the discussed evaluations focus on the comparison of computer-generated and man-made illustrations. To the best of our knowledge, no prior work has been done on quantifying the abstraction quality of computer-generated illustrations.

3 METHOD

In this section, we describe the creation of illustrations presented in our user studies. While all concepts described in the following sections can be applied to arbitrary representational drawing primitives, we illustrate them by the example of stippling. First, we introduce our normalization method that relates the number of drawing primitives to the average tonal value and size of the target image. This makes it possible to compare illustrations from different inputs. We then present the basic stippling algorithm, as well as more sophisticated adjustments required to obtain stipple drawings with differently sized points. In Section 7, we test the generalizability of our proposed approach by applying it to a different rendering primitive.

3.1 Input Normalization

The input images we have chosen are taken from typical application domains for stipple illustrations (e.g., in textbooks). Our six choices are depicted in Figure 2, where the top row shows inputs from archeology and the

Fig. 2. Study images from common domains for illustrative rendering (archaeology and biology). Top: Acinipo and Church (copyright Domingo Martín Perandrés), Lion; bottom (taken from Kim et al. (2009)): Foot, Skull, Water Flea.

bottom row from biology. When comparing stipple representations of inputs with differing characteristics it is inadequate to simply compare the number of stipple dots. To achieve a similar degree of abstraction, larger or darker images would require more points than smaller or brighter images. One obvious way to overcome this issue would be to normalize all inputs with regard to their size and brightness. However, we suggest a different form of normalization that does not require any changes to the inputs. Instead of comparing the number of stipples, we use a fraction of the summed tonal values, which we define as the inverted brightness value, from the input image as a basis for our comparisons and to deduce the needed number of points. We call this the *tonal percentage* τ. This measure is invariant to scale and content. For example, assume an image of size $1{,}000 \times 1{,}000$ pixels and an average tonal value of 0.2. Summing up each tonal value results in a tonal sum of 200k. By choosing the number of stipple dots as a tonal percentage of 10%, we would end up with 20k points for this particular image. To faithfully represent the overall tonal value of the input, the stippling algorithm has to set the point size according to the chosen number of points. Choosing a tonal percentage of 100% would create the same number of dots as produced by an error-diffusion algorithm such as Floyd–Steinberg (Floyd and Steinberg 1976).

3.2 Stippling Algorithm

Our algorithm to create stipple drawings is based on the method of Secord (2002), which makes use of a weighted variant of Lloyd's relaxation method (Lloyd 1982). It maximizes the point-to-point distances, while still maintaining a certain level of point density to represent tonal values of input images. Different variants of this algorithms are widely used, and more importantly, the number of points can be directly controlled. Other stippling algorithms (Kim et al. 2008; Li and Mould 2011; Martín et al. 2011) only offer an indirect control over this parameter. We did not adopt example-based methods, because their usage of textures would add another variable to our study. Following Secord, an initial distribution of n points $P = \{P_1, \ldots, P_n\}$ is created by rejection sampling. Then, the weighted Voronoi diagram $V = \{V_1, \ldots, V_n\}$ with Voronoi cells V_i for the initial point set P is computed. A cell V_i contains all positions $X_i = \{x \in V_i\}$ that are closer to P_i than to any other point in P (in our case,

with respect to the L2-norm). All points P_i are moved to the weighted centroids m_i of their respective cells:

$$m_i = \frac{\int_{X_i} x\rho(x)dX_i}{\int_{X_i} \rho(x)dX_i},$$

where $\rho(x)$ is a given density function, in our case the tonal value of the input image. Ultimately, the algorithm converges into a state called Centroidal Voronoi Distribution, in which all points P_i are positioned in the weighted centroids m_i of their respective Voronoi cells. Because this state is reached very slowly, the iteration is stopped when the movement of the points falls below a user-defined threshold. One limitation of this algorithm is that it can only handle points of constant size. To account for variable stipple sizes, we also employ the method proposed by Deussen et al. (2017), which shares the same principle as described before. Instead of only moving points according to their Voronoi cells, they also split and merge cells based on their size. The algorithm has two main parameters s_{min} and s_{max}, which describe the minimum and maximum size of stipples it is allowed to place. The point sizes can for example be related to local position or attributes of the input image, such as variance or the local grayscale level. In our case, we use the latter: dark areas are represented by larger points and vice versa. This does not only increase the contrast between light and dark regions making details more clearly visible but also allows spending fewer large points in dark regions, creating a more meaningful distribution of points.

3.3 Determining Point Sizes

To simplify the choices for the point size, we define $f = s_{\max}/s_{\min}$ as the factor between maximum and minimum point size, allowing us to restrict the parameter space for the stippling algorithm. We consider f fixed with $f \in \{1, 2, 3\}$. The case of $f = 1$ is special, with all points having the same size, effectively rendering the outputs equivalent to Lloyd's method. For this case, we can divide the overall tonal sum by the number of points, given by the selected tonal percentage, and deduce the actual stipples size s_{const} from the resulting area.

In case of using variable point sizes, the method has to iteratively approach the desired number of stipples, as the corresponding correct stipple sizes can not be directly set as parameter of the algorithm. To determine the correct minimum and maximum point sizes $s_{\min}$ and $s_{\max}$, we initialize both under consideration of being centered around s_{const}, which can be determined as described before. The output of this iteration will not yield the correct number of points, given by our selected tonal percentage, due to the different underlying density distribution within the input. We perform a gradient descent on the point size parameters (increase the size if there are too many points and vice versa), until the number of points is within 0.5% deviation from our target.

Figure 3 shows outputs of the algorithm with different parameter settings. The top row depicts stippling images from the same input with constant point size and increasing tonal percentages from left to right (5, 15, 25%). The center row illustrates three different inputs at the same tonal percentage (20%), resulting in different numbers of points depending on image size and content. Since the church on the right contains darker areas compared to the other two, it is represented by the highest number of points at the same tonal percentage. Finally, the bottom row shows the output when adjusting $f \in \{1, 2, 3\}$ from left to right, respectively, for the same tonal percentage and input. In Table 1, we show the radii r and number of points n for all input images used throughout this article for three normalization levels (5, 10, 15%) and three factors between minimum and maximum radius f. All input image widths were scaled down to 512 pixels to have a basis for comparison.

4 MEASURING ABSTRACTION

Several considerations impact the accuracy of assessing a subjective quantity such as the quality of a visual abstraction. When using an absolute quality scale for a single stimulus, the interpretation of scale values with respect to the subject's personal bias can be a problem: The same values might be understood differently from person to person, and the range of values can thus differ between subjects as well. An alternative approach is to present subjects two stimuli together with the task of performing a relative quality judgment. The underlying

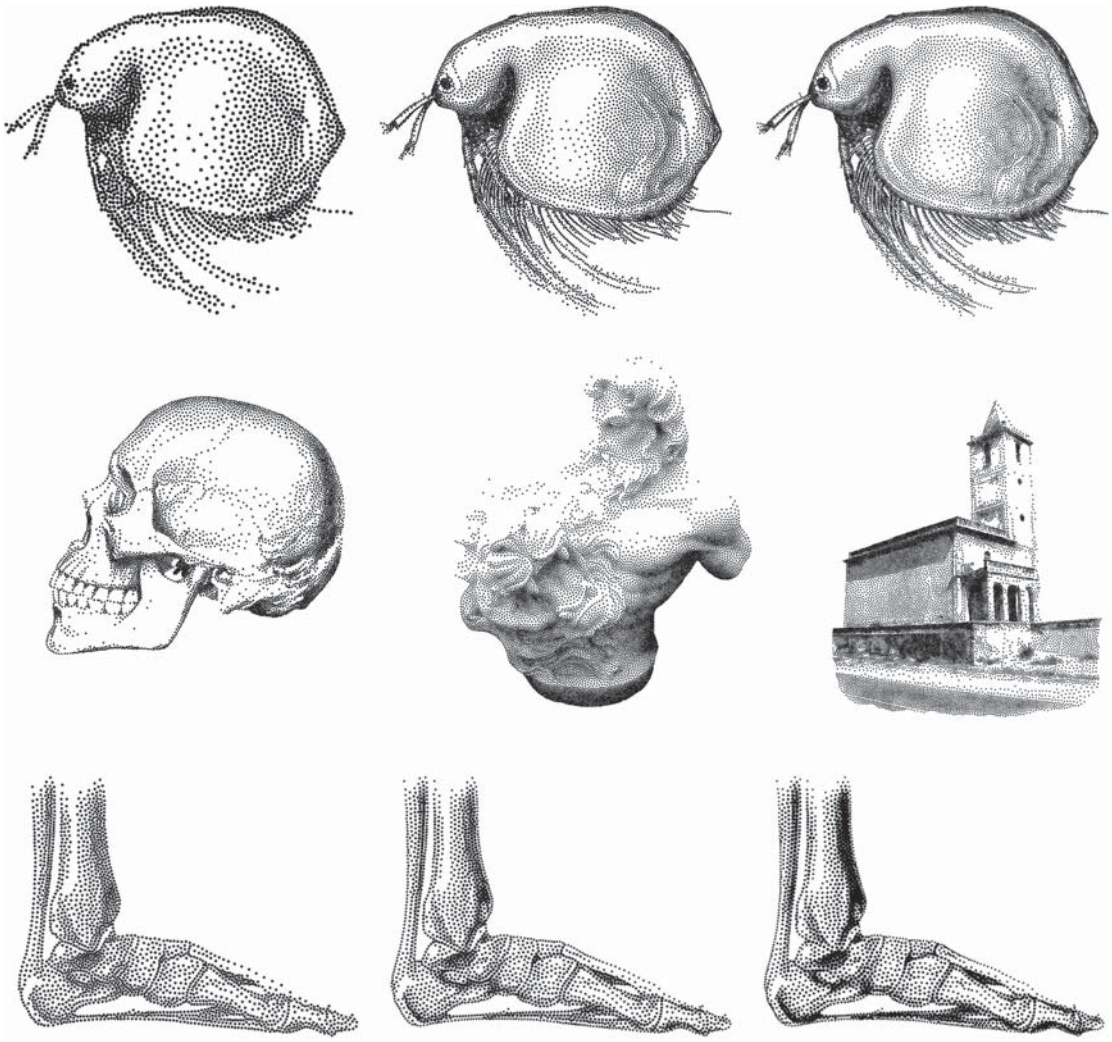

Fig. 3. Top: Stippled illustrations of the same input with increasing tonal percentages from left to right (5, 15, 25%), resulting in 1.8k, 5.5k, and 9.2k points. Center: Stippled illustrations with the same tonal percentage (20%), resulting in 4k, 9k, and 11.2k points (from left to right) due to differences in size and brightness. Bottom: Comparing the same tonal percentage (10%) and input for constant (left $f = 1$) and variable (center $f = 2$, right $f = 3$) point sizes. Using larger points in dark regions and smaller points in brighter regions increases the contrast, making details more visible.

question we are trying to answer is how the perceived quality of an abstraction is related to the number of elements used to create it. In the following, we describe how a reconstruction of absolute abstraction scale values from relative judgments can be obtained.

4.1 Thurstone's Model

Thurstone's law of comparative judgment (1927) is a well-studied method for assigning absolute scores to stimuli from relative judgment data. It has been applied in a wide variety of areas, ranging from psychology (1967) to subjective preference for video enhancement methods (2010).

Table 1. Number of Points and Radii for Three Tonal Percentages (5, 10, 15%) for All Stimuli for Stippling

Image	5%			10%			15%		
	n	r_{min}	r_{max}	n	r_{min}	r_{max}	n	r_{min}	r_{max}
				$f = 1$					
Acinipo	1,871	2.523	2.523	3,742	1.784	1.784	5,613	1.457	1.457
Church	2,807	2.523	2.523	5,614	1.784	1.784	8,421	1.457	1.457
Foot	1,068	2.523	2.523	2,136	1.784	1.784	3,204	1.457	1.457
Skull	1,013	2.523	2.523	2,026	1.784	1.784	3,039	1.457	1.457
Water Flea	1,847	2.523	2.523	3,694	1.784	1.784	5,541	1.457	1.457
Lion	2,264	2.523	2.523	4,528	1.784	1.784	6,792	1.457	1.457
				$f = 2$					
Acinipo	1,875	1.765	3.530	3,754	1.236	2.472	5,659	1.000	2.000
Church	2,825	1.789	3.578	5,639	1.247	2.493	8,445	1.009	2.018
Foot	1,064	1.971	3.941	2,129	1.377	2.753	3,204	1.121	2.242
Skull	1,012	2.000	3.999	2,028	1.390	2.780	3,047	1.119	2.237
Water Flea	1,855	1.913	3.825	3,690	1.331	2.661	5,574	1.077	2.153
Lion	2,267	1.858	3.715	4,518	1.298	2.595	6,810	1.048	2.095
				$f = 3$					
Acinipo	1,875	1.459	4.376	3,749	1.011	3.033	5,615	0.812	2.436
Church	2,820	1.472	4.415	5,645	1.010	3.030	8,468	0.810	2.430
Foot	1,071	1.703	5.108	2,131	1.187	3.562	3,184	0.958	2.873
Skull	1,017	1.782	5.345	2,016	1.226	3.677	3,034	0.985	2.955
Water Flea	1,860	1.623	4.869	3,716	1.126	3.377	5,586	0.905	2.715
Lion	2,253	1.551	4.653	4,507	1.072	3.216	6,772	0.861	2.584

Parameter f describes the factor between minimal and maximal radius.

Table 2. Coefficients and Goodness-of-Fit of the Logarithmic Regression Models for All Stimuli for Stippling

	Stimulus	α	β	γ	R^2	RMSE
			$f = 1$			
1	Acinipo	−2.2019	0.7315	−0.0544	0.9588	0.1981
2	Church	−2.2305	0.7388	−0.1234	0.9592	0.1969
3	Foot	−2.2010	0.7314	−0.0482	0.9603	0.1944
4	Skull	−2.1136	0.7095	0.1703	0.9673	0.1764
5	Water Flea	−2.1375	0.7159	0.1253	0.9703	0.1681
6	Lion	−2.2508	0.7447	−0.1506	0.9674	0.1762
	Overall	−2.1884	0.7285	−0.0086	0.9637	0.1857
			$f = 2$			
1	Acinipo	−2.0084	0.6792	0.3104	0.9322	0.2539
2	Church	−1.9523	0.6655	0.4457	0.9468	0.2250
3	Foot	−2.1667	0.7229	0.0440	0.9643	0.1844
4	Skull	−2.2006	0.7322	−0.0170	0.9713	0.1653
5	Water Flea	−2.1302	0.7144	0.1512	0.9744	0.1560
6	Lion	−2.1076	0.7077	0.1773	0.9648	0.1830
	Overall	−2.0862	0.7016	0.2083	0.9584	0.1988
			$f = 3$			
1	Acinipo	−1.8796	0.6452	0.5483	0.9351	0.2484
2	Church	−1.9718	0.6703	0.4001	0.9414	0.2361
3	Foot	−2.1504	0.7184	0.0702	0.9600	0.1952
4	Skull	−2.3113	0.7601	−0.3030	0.9685	0.1730
5	Water Flea	−2.2227	0.7376	−0.0798	0.9678	0.1751
6	Lion	−2.1299	0.7133	0.1216	0.9620	0.1902
	Overall	−2.0864	0.7013	0.1983	0.9543	0.2084

Parameter f describes the factor between minimal and maximal radius.

Thurstone's model considers judgments of a subjective quantity of stimuli $i = 1, \ldots, m$ to be a sample of Gaussian random variables $S_i \sim \mathcal{N}(\mu_i, \sigma_i^2)$, where the mean represents the true quality score. This captures the variability of responses both within a group of subjects and in repeated judgments of the same subject. A relative judgment between a pair of stimuli i and j can again be modeled as a Gaussian $S_{ij} = S_i - S_j$, or more concretely $S_{ij} \sim \mathcal{N}(\mu_{ij}, \sigma_{ij}^2)$, where $\mu_{ij} = \mu_i - \mu_j$ and $\sigma_{ij}^2 = \sigma_i^2 + \sigma_j^2$ (if S_i and S_j are uncorrelated). Thurstone's *Case V* simplifies this model further by assuming that $\sigma_i^2 = \sigma_j^2 = \frac{1}{\sqrt{2}}$, so that all $\sigma_{ij}^2 = 1$. The probability of a subject to prefer stimulus i over stimulus j is then

$$P(S_i > S_j) = P(S_i - S_j > 0) = \Phi\left(\frac{\mu_{ij}}{\sigma_{ij}}\right), \tag{1}$$

where $\Phi(\cdot)$ is the standard normal cumulative distribution function (CDF). $P(S_i > S_j)$ can be estimated from the empirical proportion of people preferring S_i over S_j. This quantity can be derived from a count matrix C, where each entry $C_{i,j}$ represents the number of times i was preferred over j. Then,

$$P(S_i > S_j) \approx \frac{C_{i,j}}{C_{i,j} + C_{j,i}}.$$

The mean quality difference μ_{ij} can be derived from inverting Equation (1), giving

$$\hat{\mu}_{ij} = \Phi^{-1}\left(\frac{C_{i,j}}{C_{i,j} + C_{j,i}}\right).$$

In this equation $\Phi^{-1}(\cdot)$ refers to the inverse standard normal CDF, or z-score. Maximum Likelihood Estimation (MLE) can be applied to estimate the scale value μ_i, $i = 1, \ldots, m$. Here, an anchoring of the values is necessary, such as $\sum \mu_i = 0$. Let μ be a vector of scale values for m stimuli $\mu = [\mu_1, \mu_2, \ldots, \mu_m]$. The log-likelihood of μ given the count matrix C can be described as

$$\mathcal{L}(\mu|C) \triangleq \log P(C|\mu) = \sum_{i,j} C_{i,j} \log(\Phi(\mu_i - \mu_j)).$$

The maximum likelihood solution scale values are obtained by solving

$$\arg\max_{\mu} \sum_{i,j} C_{i,j} \log(\Phi(\mu_i - \mu_j)) \quad \text{subj. to:} \quad \sum_i \mu_i = 0. \tag{2}$$

Tsukida and Gupta (2011) present an extensive comparison of various other techniques to estimate absolute scores, as well as detailed derivations of all the above-mentioned steps. We modified Tsukida's MATLAB implementation to non-binary scales and perform the Maximum Likelihood Estimation via the simplex search method.

4.2 Adjustments for Non-Binary Scales

The standard method for the reconstruction as described above considers a binary forced choice scale. There are two potential downsides to this approach: First, the lack of a neutral option forces participants to decide for a preference even in nearly identical pairs of stimuli. Second, participants cannot indicate different levels of granularity of their preference, meaning that strong preferences of a stimulus over any other stimulus have to be inferred from other comparisons. For this reason, we use a five-point Likert scale (Likert 1932) in our study, indicating strong and weak preferences on either side with an additional neutral option.

Consequentially, the count matrix C has to account for these different levels of preference. Let S, W, and N be count matrices, where entries $S_{i,j}$, $W_{i,j}$, and $N_{i,j}$ represent the numbers of times that the quality difference between stimuli i and j was rated strong, weak, and neutral, respectively. Further, let $\delta = \{-\delta_1, -\delta_0, \delta_0, \delta_1\}$ be the set of decision boundaries between adjacent preference options. Then, the log-likelihood of absolute scores μ given S, W, and N is defined as

$$\begin{aligned}
\mathcal{L}(\mu|S,W,N) &\triangleq \log P(S,W,N|\mu) \\
&= \sum_{i,j} S_{i,j} \log(1 - \Phi(\mu_{ij} - \delta_1)) \\
&\quad + \sum_{i,j} W_{i,j} \log(\Phi(\mu_{ij} - \delta_1) - \Phi(\mu_{ij} - \delta_0)) \\
&\quad + \sum_{i,j} N_{i,j} \log(\Phi(\mu_{ij} - \delta_0) - \Phi(\mu_{ij} + \delta_0)) \\
&\quad + \sum_{i,j} W_{j,i} \log(\Phi(\mu_{ij} + \delta_0) - \Phi(\mu_{ij} + \delta_1)) \\
&\quad + \sum_{i,j} S_{j,i} \log(\Phi(\mu_{ij} + \delta_1)).
\end{aligned}$$

Therefore, the computation of the maximum likelihood solution scale values as given in Equation (2) is augmented to

$$\arg\max_{\mu, \delta_0, \delta_1} \log P(S,W,N|\mu) \quad \text{subj. to} \quad \sum_i \mu_i = 0, \delta_1 > \delta_0 > 0. \tag{3}$$

As far as we are aware, these extensions have not been proposed in the scientific literature and could be expanded upon and generalized in the future.

5 USER STUDY

We conducted several user studies on *CrowdFlower* (now *Figure Eight*),[1] a popular crowdsourcing platform. Crowdsourcing can be understood as distributed human computing, especially useful for large sets of micro-tasks that require human intelligence. Since our paired comparisons can be understood in this way, we use this service to gather relative judgments required to obtain the perceived quality scores of different stippled abstractions. In the following, we will first discuss design considerations of our studies, followed by a description of the study design, as well as quality assurance and quality control measures employed.

5.1 Considerations for Study Design

Under consideration of what we have described in Section 3 about the different input images and choices for point size parameter f, we have to carefully design the user study so that its size does not explode, while ensuring an accurate reconstruction. Ideally, an all-to-all comparison of each input image for all choices of both f and the tonal percentage τ would yield the most accurate results. Let n be the number of f-τ combinations and k the number of stimuli, then $\mathcal{N} = k \times \frac{n(n-1)}{2}$ gives the number of paired comparisons, assuming that symmetric and identical pairs are disregarded. Consequently, with six inputs, three choices for f, and 20 samples of τ, the number of total stimuli would be $\mathcal{N} = 6 \times 1{,}770 = 10{,}620$. To reduce the required number of paired comparisons, we split the experiment in two parts: intra-model and inter-model experiments.

In intra-model experiments, we have users rate comparisons of stipple drawings with a fixed f and variable τ. An all-to-all comparison setup ensures maximum stability for the individual reconstructions and, thus, an accurate τ-quality-of-abstraction model at each level f. Each of these three studies takes $\mathcal{N}_{\text{intra}} = 6 \times 190 = 1{,}140$ paired comparisons to complete. Since each intra-model study is run separately, a reconstruction yields individually z-scored quality of abstraction scores. A direct comparison of these would not be possible. Therefore, it is necessary to determine the relative difference between scores for each factor f to make them directly comparable.

To anchor the reconstructions across different levels of f, we use inter-model experiments, where users rate comparisons of stipple drawings with a variable f and fixed τ, adding another $\mathcal{N}_{\text{inter}} = 6 \times 3 = 18$ paired comparisons for each level of τ. Tallied up, this results in $\mathcal{N}_{\text{total}} = 3{,}780$ paired comparisons to be performed, which is just about one third of the size compared to a complete all-to-all experiment. This allows the reconstruction to accurately model intra-model rankings while also ensuring inter-model comparability.

5.2 Study Design

We formulated a micro-task to be a set of three images: An original image displayed in the middle and two stippled abstractions to each side. The task of the study participants was to decide which of the two abstractions better represents the original. Options to this answers had to be given on a five-point Likert scale below the images. An example of this can be seen in Figure 1. The stipples are drawn as black anti-aliased circles.

It has to be noted here that, due to the nature of crowdsourcing, we had no influence over how the study images were displayed. The shown pairwise comparisons were scaled down to fit the participants' screen in case the resolution was too low. However, they were designed in such a way that they would fit on most screens used by participants, based on knowledge from previous experiments. We performed a pilot study using pairs of abstractions from a set of six tonal percentages $\tau_{\text{pilot}} \in \{5, 10, 15, 20, 25, 30\}$ with constant point size ($f = 1$). In this pilot, we employed an all-vs-all strategy without any restrictions. From the results, we observed a logarithmic relationship between τ and the perceived quality of an abstraction. For this reason, we sampled the 20 tonal

[1] www.crowdflower.com; now www.figure-eight.com.

percentages in our main studies logarithmically: $\tau \in \{1, 2, 3, 5, 7, 10, 13, 17, 21, 26, 31, 37, 43, 50, 57, 64, 73, 81, 90, 100\}$. Each pair of abstractions was judged by 50 individual workers. To avoid learning effects, we randomized the ordering of pairs for each worker. Additionally, we enforced a minimum distance of six image pairs before an input image was repeated.

5.3 Quality Assurance and Quality Control

Prior to performing any judgments, workers were given a short introduction into the technique of stippling, including an example illustration. In addition, users were briefly instructed on determining and rating qualities of abstractions. Study participants were given a short list of factors that may influence the judgment of an abstraction, such as reproduction of details, matching of shading, as well as the appropriateness of the number and distribution of points. Furthermore, it was stated that these factors are not an exhaustive list and that they are merely meant to give an idea of what factors can be considered when assessing the quality of an abstraction.

A quiz of ten test pairs per stimulus had to be passed by each worker, before being admitted to the main part of the study. The main body of the experiment consisted of incremental batches of ten micro-tasks per stimulus, where hidden test questions were interspersed as one of the ten micro-tasks in random fashion. Workers with an accuracy of below 70% at any point of submission had their previous answers discarded and were removed from the experiment. The ground truth for our test questions was based on the following principle:

(1) Mirco-tasks comparing identical abstractions required workers to answer with either a weak preference or the neutral option.
(2) Micro-tasks from the pilot study required workers to answer within the 90% confidence interval of the mean opinion.

Using all of these quality control measures allowed us to detect and exclude a total of 200 potentially fraudulent workers, while 892 workers from over 50 countries completed their work accurately across all experiments. Trusted workers answered an average of 296 answers, with an average test question accuracy of 92.9%. Across all comparisons, the average agreement between the judges is around 70%, where agreement is calculated as the fraction of votes for the most frequently voted preference.

6 EVALUATION

In this section, we describe the evaluation of data obtained from our user studies. Given pairwise comparison data, we reconstruct absolute abstraction scores using the approach described in Section 4. We propose a logarithmic regression model, fit it to our data, and show the accuracy of the model on a different dataset. First, we describe the process for stipple drawings with constant point size (intra-model), and later on, we show how to make stimuli from different algorithms or parameter settings comparable (inter-model) with a second form of study.

6.1 Voting Behavior

Crowd workers could describe the quality differences between any given pair of abstractions by three orders of magnitude (neutral, weak, strong). The number of votes for each of these preferential options (averaged over all input images) for the abstractions created with a constant point size are displayed in Figure 4. For the purpose of visualizing the voting behavior, we sorted the pairwise comparisons in the count matrices so that the abstraction with lower tonal percentage is on the left.

In the left image, we can see that the neutral option was mainly chosen for abstraction pairs with the same tonal percentages. With increasing tonal percentages of both stimuli, the differences between two abstract representations become more difficult to identify, because visual differences become more subtle. Therefore, the neutral option was also more often chosen for high but unequal tonal percentages, which can be seen by the expanding region in which at least 20 out of 50 possible votes were cast, outlined in red. The same pattern can

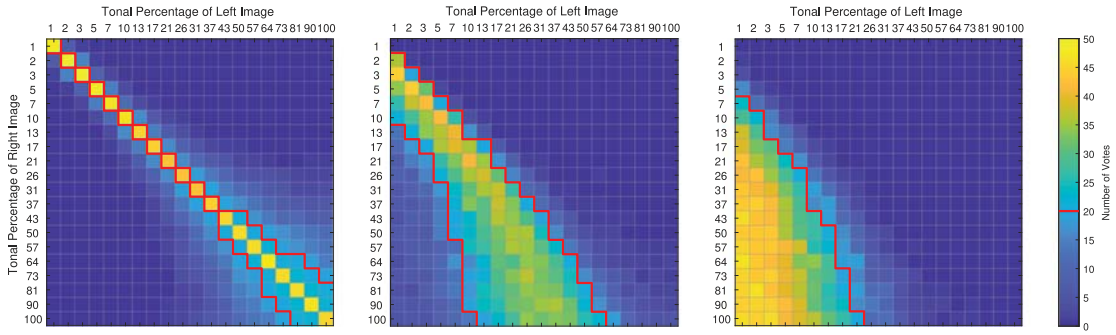

Fig. 4. Visualization of the voting behavior for abstractions created with constant point size. Left to right: count matrices for neutral (N), weak (W), and strong (S) preference options on the five-point Likert scale for all 20 tonal percentages averaged over all stimuli. The red line indicates where at least 20 out of the 50 possible votes were cast. For visualization purposes the abstraction pairs were sorted so that the abstraction with lower tonal percentage is on the left.

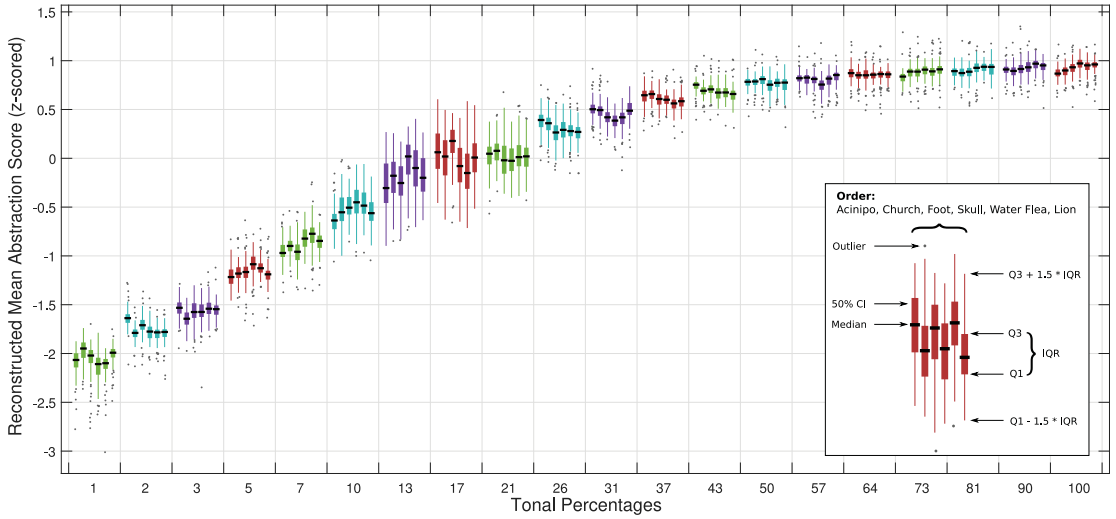

Fig. 5. Reconstructed absolute abstraction scores for 100 reconstructions of all stimuli as boxplots grouped by tonal percentage. The black bar indicates the median abstraction score and the body represents the 50% confidence interval (IQR). The whiskers depict the range from lower/upper quartiles to the last data point within 1.5 · IQR and outliers are shown as gray dots.

also be found for the weak and strong preferential options in the middle and right image. As expected, the weak preference was chosen for smaller tonal differences, whereas the strong preference was predominantly chosen for larger ones.

6.2 Reconstruction of Abstraction Scores

We can now compute absolute mean abstraction scores (MAS) by applying the reconstruction algorithm proposed in Equation (3). In Figure 5, we show the result of this reconstruction for all six input images, using a constant point size. Since the reconstructed results are not necessarily exactly the same due to random initialization of the boundaries between options on the Likert scale, we show distributions of 100 such reconstructions

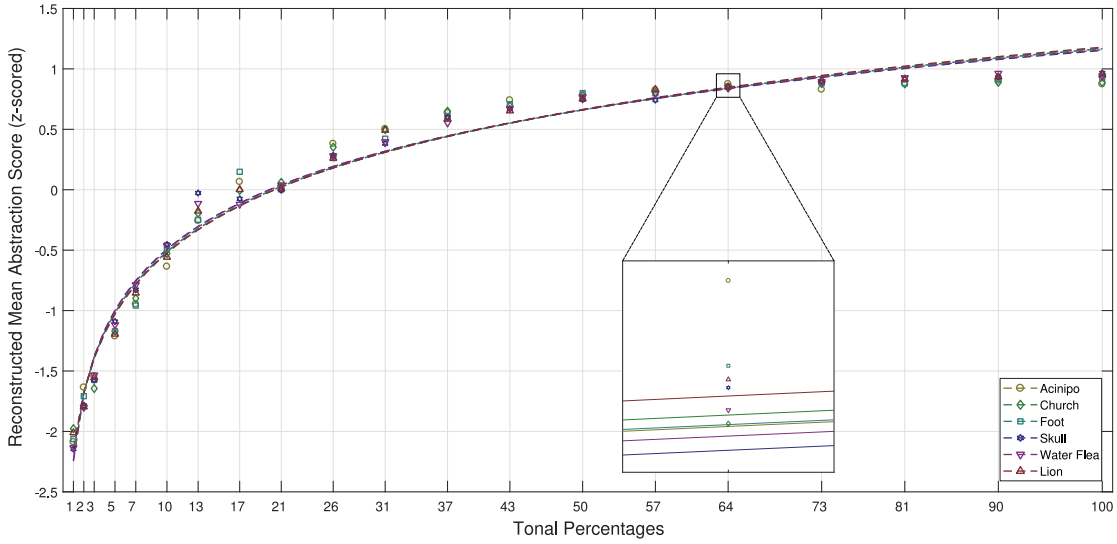

Fig. 6. Fitted logarithmic models for all six stimuli alongside the mean abstraction scores averaged over 100 reconstructions using Thurstonian scaling for stippling with constant point size. The models for each different stimulus are almost identical, indicating that the perceived quality is independent of the content of the input.

as boxplots, grouped by tonal percentage. In a next step, we want to create a model to predict the abstraction score from a given tonal percentage. Because of the observed logarithmic relationship between tonal percentage and MAS in this visualization, we propose a logarithmic regression model with the closed form formula:

$$\mathrm{MAS}(\tau) = \alpha + \beta \cdot \ln(\tau - \gamma),$$

where τ is the tonal percentage, and α, β, and γ are stimuli specific coefficients. The fitted curves for this model alongside the mean abstraction score of the 100 reconstructions are depicted in Figure 6. The individual model coefficients and goodness-of-fit statistics are denoted in Table 2.

The resulting overall model $\mathrm{MAS}(\tau) = -2.1884 + 0.7285 \cdot \ln(\tau + 0.0086)$ is in accordance with *Weber–Fechner's Law (WFL)* (Fechner 1860), which describes a logarithmic relationship between human perception and physical stimuli. The just noticeable threshold of change in stimulus intensity to the intensity of the original stimulus was described by Ernst Heinrich Weber in 1834 as a constant. Gustav Fechner later formulated that the differential perception is proportional to the relative change of the stimulus:

$$d\mathcal{P} = k \cdot \frac{dI}{I},$$

with I being the stimulus intensity, and k a sense-specific constant. Integration of this equation leads to

$$\mathcal{P} = k \cdot \ln \frac{I}{I_0}.$$

Here, $\mathcal{P}$ is the perceptual intensity and I_0 is a constant introduced through integration that can be interpreted as a stimulus-specific perceptional threshold. The WFL has been shown to hold for a wide range of perceptual scenarios, one of which being human vision. For a more in-depth discussion of the WFL in the area of perception, we refer to the work of Reichl et al. (2010). A novel finding here is that the WFL seems to also be applicable to the number of drawing primitives (stipples) in abstract illustrations. Furthermore, since the reconstructed models for each stimulus are almost identical, this seems to be independent of the content of the input. While it was

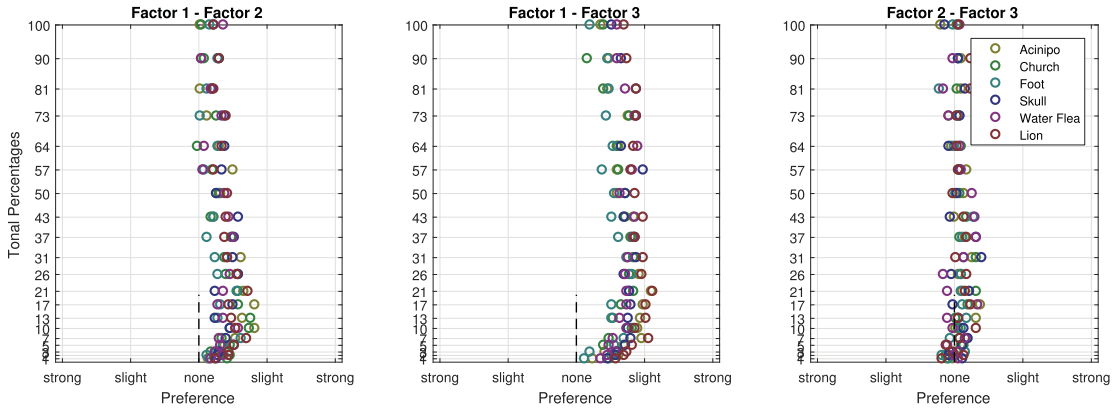

Fig. 7. Direct comparison of stipple drawings created with constant (factor 1) and variable point size (factors 2 and 3). The preference toward either is visualized as horizontal position. Shown is a preference toward adaptive point sizes, which becomes weaker toward higher tonal percentages.

already discovered that we see arrangements of objects as a whole by King and Wertheimer (2004), it was not apparent that this could also apply for the perceived quality of abstractions.

6.3 Abstraction Scores between Models

To make models from different algorithms or parameter settings directly comparable they have to share a common scale. For this purpose, we conducted a second form of user studies that compared results created with different settings of f for the same tonal percentage against each other. The same five-point Likert rating scale was used as in the other study format. The preferential options are displayed in Figure 7. While a general preference toward the variable point size models can be observed, this tendency becomes weaker toward higher tonal percentages. The same effect has been observed when comparing similar high tonal percentages for the same model, where it became more difficult to compare representations with a high amount of elements and therefore complexity. Study participants also showed a neutral preference at very low tonal percentages, which can be explained by the low quality of both abstract representations with very few rendering primitives. Since both variable point size models for $f = 2$ and $f = 3$ increase the contrast, the difference between the two is smaller compared to the constant point size model $f = 1$.

Having determined differences from the relative comparison with different factors f, we can now scale and shift the scores from each individual model so that a direct comparison is possible by using the same scale for all. The adapted reconstructed models can be seen as dashed black lines in Figure 8. The filled areas around them depict the 95% confidence interval of each model, and the scattered points indicate results from a confirmation study, which will be described in the next section. We expected that by using a more intelligent point distribution approach, the absolute abstraction scores could already be higher at lower tonal percentages. By using variable point sizes depending on the local grayscale value, we can save a large number of points in darker regions by simply increasing their size, thereby not only boosting the visual contrast but also allowing better placement of the saved points. When looking at any of the variable point size models, we can see that this assumption holds. The scores increase more rapidly compared to the model for constant point size, while the differences become smaller at higher tonal percentages. This latter observation can be explained by the level of detail retained through the large number of small points at higher tonal percentages, which makes differentiation more difficult.

To better understand the implications of changing factor f with respect to the resulting abstraction scores, we propose a quantification in terms of primitive-savings similar to the Bjøntegaard-Delta bitrate (Bjontegaard

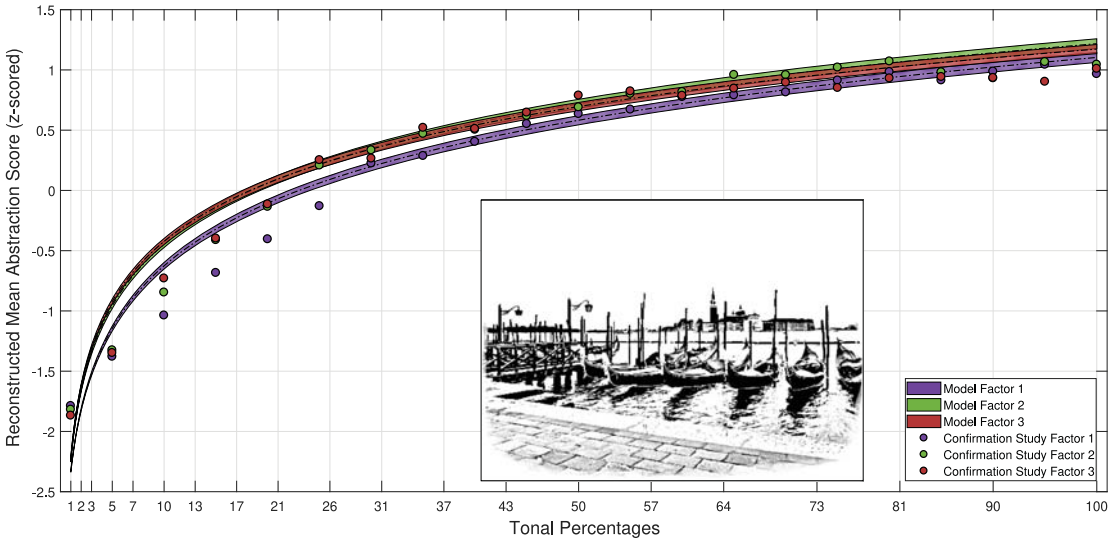

Fig. 8. Comparisons of models for different stippling settings of f sharing a common scale. The dashed black lines indicate the models' means, the colored area around it depict the 95% confidence intervals. Dots are abstraction scores for the evaluation study on a different stimulus, shown as inset. Image copyright Domingo Martín Perandrés.

2001). Assume we are given an image to create a stippled representation of: With the choice of $f \in \{1, 3\}$ it can be estimated that by increasing the factor f, we can reduce τ while retaining the abstraction quality. This reduction in tonal percentage directly translates into a reduction of primitives used. On average, the sum of τ-differences at the same abstraction score normalized according to the tonal percentage gives us the primitive-savings. Between factors $f = 1$ and $f = 3$, a 17.5% reduction of τ is achieved on average at the same level of abstraction, implying a 17.5% reduction of primitives required. When comparing the case of $f = 1$ and $f = 2$, the average primitives-saving is around 19%. This shows that the choice for a smarter distribution of points by adjusting the point size allows for a significant reduction in stipple count.

6.4 Accuracy of Prediction

We evaluate the performance and verify the generality of our reconstructed model by conducting an additional evaluation study with a different input image for each point size setting. This experiment follows the same settings as our main study (cf. Section 5), except that the participants were different and the set of tonal percentages was selected at different levels to assure the generalizability of our approach. The evaluation tonal percentages were chosen to be more equally distributed: $\tau_{\text{eval}} \in \{1, 5, 10, 15, \ldots, 95, 100\}$. Let $\hat{\mu}_{\text{eval}}$ be the absolute MAS scores obtained with our previously described approach. Then, the distribution of these scores has to be fit toward the expected distribution of our model by adjusting them to have the same population mean and standard deviation as $\text{MAS}(\tau_{\text{eval}})$. Figure 8 shows these fitted confirmation scores for each model, presenting a good fit between our model and the evaluation data: The models explain 93.99%, 92.89%, and 92.21% of the variability of the data for $f \in \{1, 2, 3\}$, respectively. The RMSEs are also relatively low at 0.3027, 0.4085, and 0.3959, albeit higher than for the original sets of stimuli at logarithmically sampled τ values (cf. Table 2).

7 BEYOND POINT PRIMITIVES: HATCHING

To show the generalizability of our proposed approach, we also apply it to a different stylization method. In hatching, closely spaced parallel lines are used to create tonal effects, with crossing lines being used to represent

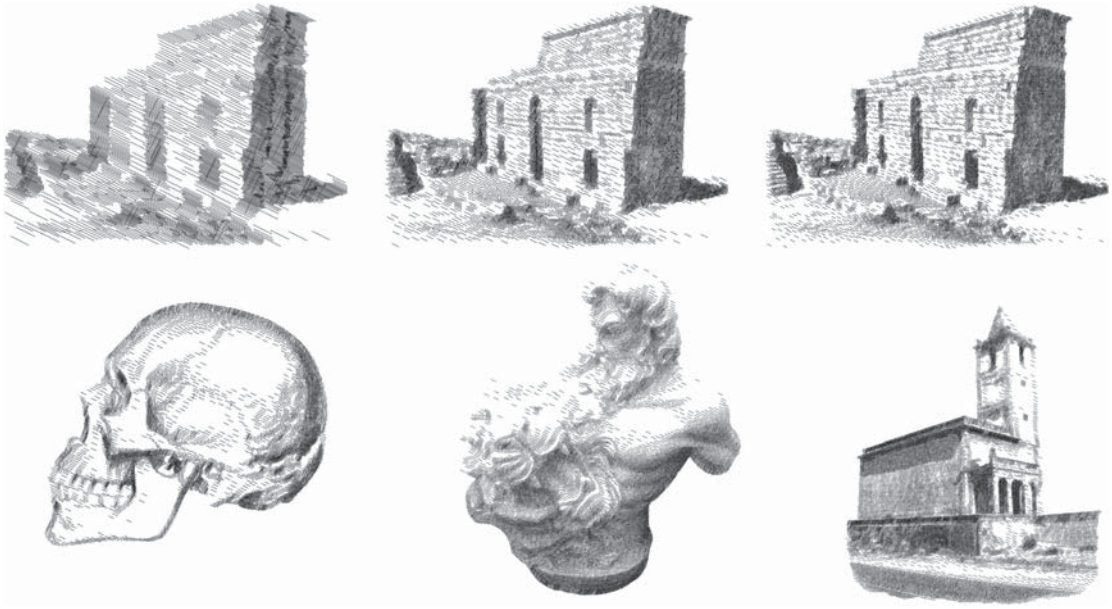

Fig. 9. Top: Hatching results created from the same input with increasing tonal percentages (5, 15, 25%). Bottom: Hatching results created with the same tonal percentage (20%). The total number of lines is in both cases exactly the same as the number of points of the respective results in Figure 3.

dark regions (crosshatches). Because lines have more parameters than points, we had to restrict our technique to keep the parameter space, and thereby the user study size, manageable. Our proposed method uses lines with fixed orientations, both for hatches and crosshatches, and a fixed line width, which is common in these forms of illustrations due to a fixed toolset. We only varied the number of lines and their length according to the tonal percentage measure described before. The same tonal percentage approach was employed as described for stippling with variable point size to normalize the number of lines. For increasing tonal percentages, we decreased the line length l to increase their number. We set the length to $l = L/\tau$, with L being a fixed parameter (in our case 100 pixels). The line orientations were fixed to $\pi/4$ for hatches and perpendicular with $-\pi/4$ for crosshatches.

Our hatching results were generated with the algorithm of Deussen et al. (2017), which we also used for stippling with variable point sizes. While the algorithm is originally intended to be used for stippling, only the Voronoi diagram calculation has to be adapted for lines. The splitting strategy and relaxation were kept the same as for stippling, except that we placed a line centered at the stipple position. We used the algorithm twice, first to create hatches for brighter regions, and once for crosshatches in darker regions. For the first pass, we clamped the input brightness values below a threshold, and for the second pass, we only used values that are below said threshold. The threshold has been defined per input and was in a range between 0.3 and 0.5. The resulting lines from both steps were then merged to create the final result. We show results from our hatching algorithm in Figure 9. The same input with an increasing tonal percentage is depicted at the top of Figure 9 for 5, 15, and 25%. The results of different inputs for the same tonal percentage are shown at the bottom. Since we use the same tonal percentage approach, the number of lines are the same as points for stippling, and the darkest image is represented with the most lines.

The reconstructed absolute abstraction score and the fitted model for our hatching approach is shown in Figure 10. When comparing it to the model for stippling in Figure 6, it can be seen that both are quite similar in

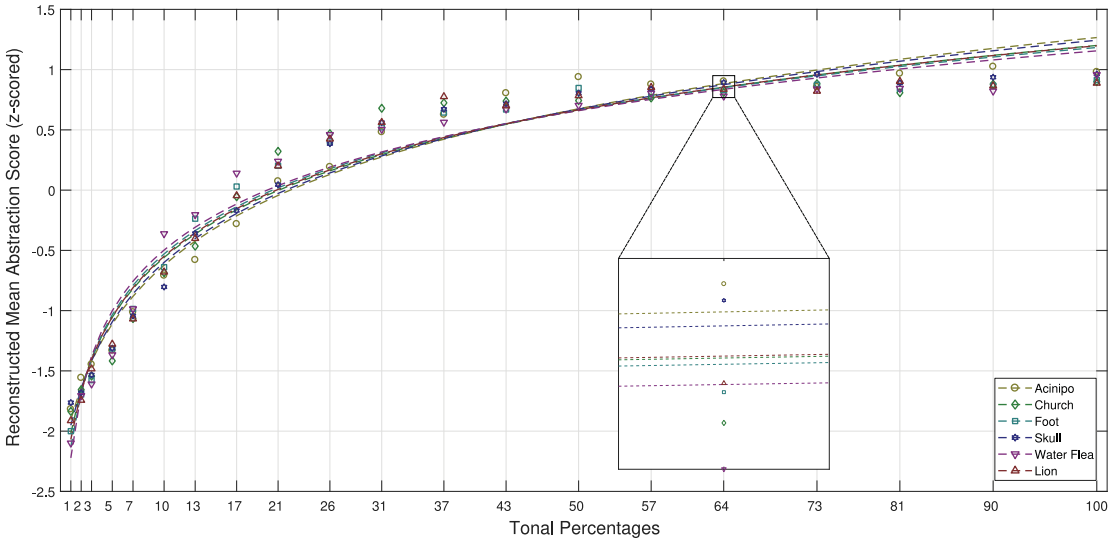

Fig. 10. Fitted logarithmic models for all six stimuli alongside the mean abstraction scores for hatching. While the models are similar to stippling (cf. Figure 6), there is more variance between the different stimuli, indicating a higher dependence of hatching on the input.

shape, although the absolute values are not directly comparable. One difference is that the scores for individual inputs, and therefore the different models for each input, have a higher variance compared to stippling. We assume one reason for this is that the quality of an abstraction using lines is more dependent on the input compared to using points, i.e. an input containing many curved forms can still be easily represented with a set of dots, but less so with straight lines.

8 CONCLUSION AND FUTURE WORK

We investigated how the perceived abstraction quality of computer-generated illustrations is related to the number of drawing primitives used to create them. For this, we presented a novel approach to derive perceptual models for such illustrations from a user study by gathering comparative data and employing a paired comparison model. To compare illustrations from different inputs, we introduced the scale and content invariant tonal percentage measure. Using this approach, we conducted an exemplar study investigating the relation between number of points and abstraction quality of stipple drawings, showing a logarithmic dependency independent of the input image. This can be related to Weber–Fechner's law from psychophysics, which states that the relationship between a stimulus and its perception is logarithmic. Furthermore, we investigated the impact of adjusting the size of points used in the stippling process. By showing that our proposed approach also works for a different output primitive for small lines, we demonstrate its applicability toward other rendering primitives. It also suggests that the logarithmic relationship between number of rendering primitives and perceived abstraction quality, and therefore Weber–Fechner's law, might hold for abstract rendering primitives in general.

Our results can help users to choose an adequate number of drawing primitives in computer-generated illustrations for the presented stippling and hatching techniques. Usually, the number of elements is a user-defined parameter and there is a trade-off between quality and computation time. Since our model of abstraction quality is steadily increasing for stippling and hatching, it is not possible to derive an optimal value. For interpretation of these values, we suggest to look at the presented function's gradient and decide whether the increment in number of elements, and therefore computation time, is worth the corresponding change in quality.

While we presented a general approach to derive perceptual models for a specific form of illustrations, there are limitations to its applicability and our presented results. First, it is difficult and beyond the scope of this article to prove that our approach holds for every conceivable drawing element and creates consistent results with ours. By comparing two different methods for stippling and one for hatching, we have at least an indication that this assumption might hold. However, we have provided the necessary tools to investigate this further in future works. It is important to note that comparing the number of distinct rendering elements to the perceived quality of the respective illustration is only meaningful in the absence of overlaps. Otherwise, it would be possible to add occluded elements, which do not contribute to the perceived quality, contradicting our model. This is especially problematic for more spacious drawing elements, such as textured strokes for painterly rendering, limiting our approach to methods that avoid or at least limit the amount of overlap between rendering elements. Furthermore, we currently do not include semantics into our model and treat the image with uniform importance.

While an additional study with three portraits showed that our models for stippling were similar to the ones presented in this article, we expect this to differ when treating certain regions with more importance: For stippling, more points could be distributed in the eyes or mouth areas. These regions are known to play a decisive role in face-processing, which is performed by specialized parts of the visual cortex (Johnson 2005; Kanwisher et al. 1997; Tsao et al. 2006). Here, we expect higher abstraction scores for the same number of points compared to our current models.

For future work, salient regions that have a stronger impact on the perceived quality of an abstraction could be studied separately to improve the model, e.g., by eye tracking methodologies or equivalent crowdsourced approaches (Engelke and Le Callet 2015; Hosu et al. 2016; Jiang et al. 2015). Conducting similar studies to the ones presented in this article for additional illustration techniques, including those combining different drawing elements, would be a compelling topic for further investigations. Last, the equivalence of two illustrations is only drawn analytically by a graph of preference ratings. Perceptual anomalies may, however, cause parts of the image to be more or less salient in one abstraction, especially for different stipple distribution strategies. In turn, the way abstractions are observed by viewers could differ. Comparing scanpaths obtained in an eye tracking experiment could further support whether two different abstractions with the same quality are to be considered equal or not. Additionally, it would be interesting to see if guiding the point allocation strategy according to saliency could further reduce the number of required primitives for the same abstraction quality.

REFERENCES

Michael Balzer, Thomas Schlömer, and Oliver Deussen. 2009. Capacity-constrained Point Distributions: A variant of Lloyd's method. *ACM Trans. Graph.* 28, 3, Article 86 (July 2009), 8 pages. DOI: https://doi.org/10.1145/1531326.1531392

Pascal Barla, Simon Breslav, Joëlle Thollot, François Sillion, and Lee Markosian. 2006. Stroke pattern analysis and synthesis. *Comput. Graph. Forum* 25, 3 (2006), 663–671. DOI: https://doi.org/10.1111/j.1467-8659.2006.00986.x

Gisle Bjontegaard. 2001. Calculation of average PSNR differences between RD-curves. In *Proceedings of the 13th Meeting of the Video Experts Group (VCEG'01).*

Forrester Cole, Kevin Sanik, Doug DeCarlo, Adam Finkelstein, Thomas Funkhouser, Szymon Rusinkiewicz, and Manish Singh. 2009. How well do line drawings depict shape? *ACM Trans. Graph.* 28, 3, Article 28 (July 2009). DOI: https://doi.org/10.1145/1531326.1531334

Clyde H. Coombs. 1967. Thurstone's measurement of social values revisited forty years later. *J. Personal. Soc. Psychol.* 6, 1 (1967), 85. DOI: https://doi.org/10.1037/h0024522

Fernando de Goes, Katherine Breeden, Victor Ostromoukhov, and Mathieu Desbrun. 2012. Blue noise through optimal transport. *ACM Trans. Graph.* 31, 6, Article 171 (Nov. 2012). DOI: https://doi.org/10.1145/2366145.2366190

Oliver Deussen. 2009. Aesthetic placement of points using generalized Lloyd relaxation. In *Computational Aesthetics in Graphics, Visualization, and Imaging*. The Eurographics Association. DOI: https://doi.org/10.2312/COMPAESTH/COMPAESTH09/123-128

Oliver Deussen, Stefan Hiller, Cornelius Van Overveld, and Thomas Strothotte. 2000. Floating points: A method for computing stipple drawings. *Comput. Graph. Forum* 19, 3 (2000), 41–50. DOI: https://doi.org/10.1111/1467-8659.00396

Oliver Deussen and Tobias Isenberg. 2013. *Halftoning and Stippling*. Springer, 45–61. DOI: https://doi.org/10.1007/978-1-4471-4519-6_3

Oliver Deussen, Marc Spicker, and Qian Zheng. 2017. Weighted Linde–Buzo–Gray stippling. *ACM Trans. Graph.* 36, 6, Article 233 (2017). DOI: https://doi.org/10.1145/3130800.3130819

Frédo Durand, Victor Ostromoukhov, Mathieu Miller, Julie Duranleau, and François'and Dorsey. 2001. Decoupling strokes and high-level attributes for interactive traditional drawing. In *Rendering Techniques 2001*. Springer, Vienna, 71–82.

Ulrich Engelke and Patrick Le Callet. 2015. Perceived interest and overt visual attention in natural images. *Signal Process.: Image Commun.* 39 (2015), 386–404.

Gustav T. Fechner. 1860. *Elemente der Psychophysik*, Vol. 1–2. Breitkopf & Härtel.

R. W. Floyd and L. Steinberg. 1976. An adaptive algorithm for spatial grey scale. In *Proceedings of the Society of Information Display*. 75–77.

C. Gatzidis, S. Papakonstantinou, V. Brujic-Okretic, and S. Baker. 2008. Recent advances in the user evaluation methods and studies of non-photorealistic visualization and rendering techniques. In *Information Visualisation*. IEEE, 475–480. DOI: https://doi.org/10.1109/IV.2008.75

Stéphane Grabli, Emmanuel Turquin, Frédo Durand, and François X. Sillion. 2010. Programmable rendering of line drawing from 3D scenes. *ACM Trans. Graph.* 29, 2, Article 18 (2010), 20 pages. DOI: https://doi.org/10.1145/1731047.1731056

L. Harrison, F. Yang, S. Franconeri, and R. Chang. 2014. Ranking visualizations of correlation using weber's law. *IEEE Trans. Visual. Comput. Graph.* 20, 12 (Dec. 2014), 1943–1952. DOI: https://doi.org/10.1109/TVCG.2014.2346979

Stefan Hiller, Heino Hellwig, and Oliver Deussen. 2003. Beyond stippling—Methods for distributing objects on the plane. *Comput. Graph. Forum* 22, 3 (2003), 515–522. DOI: https://doi.org/10.1111/1467-8659.00699

Elaine R. S. Hodges. 2003. *The Guild Handbook of Scientific Illustration*, 2nd ed. Wiley.

Vlad Hosu, Franz Hahn, Igor Zingman, and Dietmar Saupe. 2016. Reported attention as a promising alternative to gaze in IQA tasks. In *Proceedings of the 5th ISCA/DEGA Workshop on Perceptual Quality of Systems (PQS'16)*. 117–121.

Tobias Isenberg. 2013. *Evaluating and Validating Non-photorealistic and Illustrative Rendering*. Springer, London, 311–331. DOI: https://doi.org/10.1007/978-1-4471-4519-6_15

Tobias Isenberg, Petra Neumann, Sheelagh Carpendale, Mario Costa Sousa, and Joaquim A. Jorge. 2006. Non-photorealistic rendering in context: An observational study. In *Proceedings of the 4th International Symposium on Non-photorealistic Animation and Rendering (NPAR'06)*. ACM, New York, NY, 115–126. DOI: https://doi.org/10.1145/1124728.1124747

Ming Jiang, Shengsheng Huang, Juanyong Duan, and Qi Zhao. 2015. Salicon: Saliency in context. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR'15)*. IEEE, 1072–1080.

Pierre-Marc Jodoin, Emric Epstein, Martin Granger-Piché, and Victor Ostromoukhov. 2002. Hatching by example: A statistical approach. In *Proceedings of the 2nd International Symposium on Non-photorealistic Animation and Rendering (NPAR'02)*. ACM, New York, NY, 29–36. DOI: https://doi.org/10.1145/508530.508536

Mark H. Johnson. 2005. Subcortical face processing. *Nature Rev. Neurosci.* 6, 10 (2005), 766–774. DOI: https://doi.org/10.1038/nrn1766

Nancy Kanwisher, Josh McDermott, and Marvin M. Chun. 1997. The fusiform face area: A module in human extrastriate cortex specialized for face perception. *J. Neurosci.* 17, 11 (1997), 4302–4311. http://www.jneurosci.org/content/17/11/4302.

M. Kay and J. Heer. 2016. Beyond weber's law: A second look at ranking visualizations of correlation. *IEEE Trans. Visual. Comput. Graph.* 22, 1 (Jan. 2016), 469–478. DOI: https://doi.org/10.1109/TVCG.2015.2467671

Dongyeon Kim, Minjung Son, Yunjin Lee, Henry Kang, and Seungyong Lee. 2008. Feature-guided image stippling. In *Proceedings of the 19th Eurographics Conference on Rendering (EGSR'08)*. 1209–1216. DOI: https://doi.org/10.1111/j.1467-8659.2008.01259.x

Sung Ye Kim, Ross Maciejewski, Tobias Isenberg, William M. Andrews, Wei Chen, Mario Costa Sousa, and David S. Ebert. 2009. Stippling by example. In *Proceedings of the 7th International Symposium on Non-Photorealistic Animation and Rendering (NPAR'09)*. ACM, New York, NY, 41–50. DOI: https://doi.org/10.1145/1572614.1572622

D. Brett King and Michael Wertheimer. 2004. *Max Wertheimer and Gestalt Theory*. Transaction Publishers.

Johannes Kopf, Daniel Cohen-Or, Oliver Deussen, and Dani Lischinski. 2006. Recursive wang tiles for real-time blue noise. In *Proceedings of the ACM SIGGRAPH 2006 Papers (SIGGRAPH'06)*. ACM, New York, NY, 509–518. DOI: https://doi.org/10.1145/1179352.1141916

Hua Li and David Mould. 2011. Structure-preserving stippling by priority-based error diffusion. In *Proceedings of Graphics Interface (GI'11)*. Canadian Human-Computer Communications Society, 127–134. Retrieved from http://dl.acm.org/citation.cfm?id=1992917.1992938.

Rensis Likert. 1932. A technique for the measurement of attitudes. *Arch. Psychol.* 22, 140 (1932), 1–55.

Stuart Lloyd. 1982. Least squares quantization in PCM. *IEEE Trans. Info. Theory* 28, 2 (1982), 129–137. DOI: https://doi.org/10.1109/TIT.1982.1056489

Ross Maciejewski, Tobias Isenberg, William M. Andrews, David S. Ebert, and Mario Costa Sousa. 2007. Aesthetics of hand-drawn vs. computer-generated stippling. In *Proceedings of the 3rd Eurographics Conference on Computational Aesthetics in Graphics, Visualization and Imaging (Computational Aesthetics'07)*. Aire-la-Ville, Switzerland, 53–56. DOI: https://doi.org/10.2312/COMPAESTH/COMPAESTH07/053-056

R. Maciejewski, T. Isenberg, W. M. Andrews, D. S. Ebert, M. C. Sousa, and W. Chen. 2008. Measuring stipple aesthetics in hand-drawn and computer-generated images. *IEEE Comput. Graph. Appl.* 28, 2 (Mar. 2008), 62–74. DOI: https://doi.org/10.1109/MCG.2008.35

Domingo Martín, Germán Arroyo, M. Victoria Luzón, and Tobias Isenberg. 2011. Scale-dependent and example-based grayscale stippling. *Comput. Graph.* 35, 1 (2011), 160–174. DOI: https://doi.org/10.1016/j.cag.2010.11.006

Domingo Martín, Vicente del Sol, Celia Romo, and Tobias Isenberg. 2015. Drawing characteristics for reproducing traditional hand-made stippling. In *Proceedings of the Workshop on Non-Photorealistic Animation and Rendering (NPAR'15)*. Eurographics Association, Aire-la-Ville, Switzerland, 103–115. Retrieved from http://dl.acm.org/citation.cfm?id=2810002.2810007.

Domingo Martín, Germán Arroyo, Alejandro Rodríguez, and Tobias Isenberg. 2017. A survey of digital stippling. *Comput. Graph.* 67 (2017), 24–44. DOI: https://doi.org/10.1016/j.cag.2017.05.001

David Mould. 2007. Stipple placement using distance in a weighted graph. In *Proceedings of the 3rd Eurographics Conference on Computational Aesthetics in Graphics, Visualization and Imaging (Computational Aesthetics'07)*. Eurographics Association, Aire-la-Ville, Switzerland, 45–52. DOI: https://doi.org/10.2312/COMPAESTH/COMPAESTH07/045-052

Oscar M. Pastor, Bert Freudenberg, and Thomas Strothotte. 2003. Real-time animated stippling. *IEEE Comput. Graph. Appl.* 23, 4 (2003), 62–68. DOI: https://doi.org/10.1109/MCG.2003.1210866

P. Reichl, S. Egger, R. Schatz, and A. D'Alconzo. 2010. The logarithmic nature of QoE and the role of the Weber-Fechner law in QoE assessment. In *Proceedings of the IEEE International Conference on Communications*. 1–5. DOI: https://doi.org/10.1109/ICC.2010.5501894

Vera Rivotti, João Proença, Joaquim Jorge, and Mário Costa Sousa. 2007. Composition principles for quality depiction and aesthetics. In *Proceedings of the 3rd Eurographics Conference on Computational Aesthetics in Graphics, Visualization and Imaging (Computational Aesthetics'07)*. Eurographics Association, Aire-la-Ville, Switzerland, 37–44. DOI: https://doi.org/10.2312/COMPAESTH/COMPAESTH07/037-044

Adrian Secord. 2002. Weighted voronoi stippling. In *Proceedings of the 2nd International Symposium on Non-photorealistic Animation and Rendering (NPAR'02)*. ACM, New York, NY, 37–43. DOI: https://doi.org/10.1145/508530.508537

Mayank Singh and Scott Schaefer. 2010. Suggestive hatching. In *Computational Aesthetics in Graphics, Visualization, and Imaging*, Pauline Jepp and Oliver Deussen (Eds.). The Eurographics Association. DOI: https://doi.org/10.2312/COMPAESTH/COMPAESTH10/025-032

Marc Spicker, Franz Hahn, Thomas Lindemeier, Dietmar Saupe, and Oliver Deussen. 2017. Quantifying visual abstraction quality for stipple drawings. In *Proceedings of the Symposium on Non-Photorealistic Animation and Rendering (NPAR'17)*. ACM, New York, NY, 8:1–8:10. DOI: https://doi.org/10.1145/3092919.3092923

Louis L. Thurstone. 1927. A law of comparative judgment. *Psychol. Rev.* 34, 4 (1927), 273. DOI: https://doi.org/10.1037/h0070288

Doris Y. Tsao, Winrich A. Freiwald, Roger B. H. Tootell, and Margaret S. Livingstone. 2006. A cortical region consisting entirely of face-selective cells. *Science* 311, 5761 (2006), 670–674. DOI: https://doi.org/10.1126/science.1119983

Kristi Tsukida and Maya R. Gupta. 2011. *How to Analyze Paired Comparison Data*. Technical Report. DTIC Document.

Georges Winkenbach and David H. Salesin. 1994. Computer-generated pen-and-ink illustration. In *Proceedings of the 21st Annual Conference on Computer Graphics and Interactive Techniques (SIGGRAPH'94)*. ACM, New York, NY, 91–100. DOI: https://doi.org/10.1145/192161.192184

Russell L. Woods, PremNandhini Satgunam, P. Matthew Bronstad, and Eli Peli. 2010. Statistical analysis of subjective preferences for video enhancement. *Human Vision and Electronic Imaging XV* 7527 (2010). DOI: https://doi.org/10.1117/12.843858