Quantifying Visual Abstraction Quality for Stipple Drawings

Marc Spicker
University of Konstanz
Konstanz, Germany

Franz Hahn
University of Konstanz
Konstanz, Germany

Thomas Lindemeier
University of Konstanz
Konstanz, Germany

Dietmar Saupe
University of Konstanz
Konstanz, Germany

Oliver Deussen
University of Konstanz
Konstanz, Germany

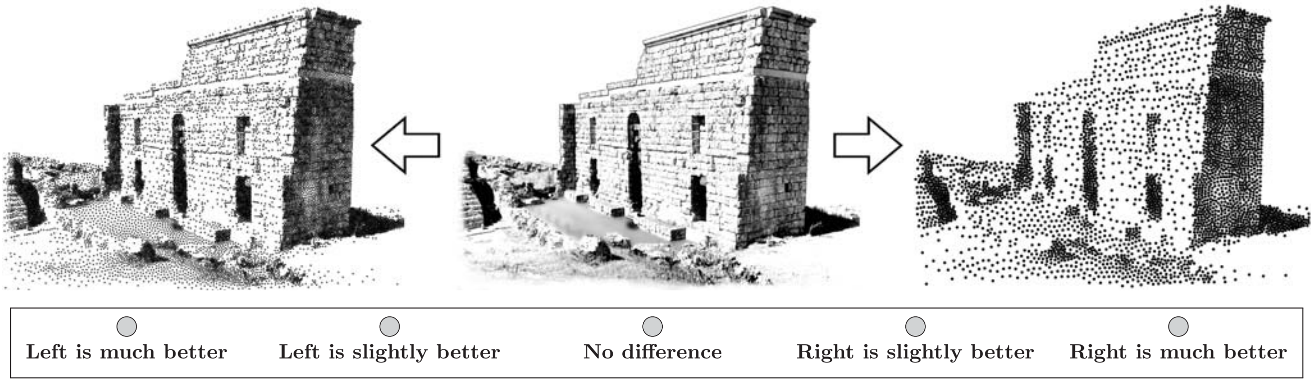

Figure 1: Comparing the abstraction quality of two stipple illustrations (left ~15k and right ~4k points) to an input image taken from our user study. We use this relative information to derive values for the absolute perceived abstraction quality of stippled representations.

ABSTRACT

We investigate how the perceived abstraction quality of stipple illustrations is related to the number of points used to create them. Since it is difficult to find objective functions that quantify the visual quality of such illustrations, we gather comparative data by a crowdsourcing user study and employ a paired comparison model to deduce absolute quality values. Based on this study we show that it is possible to predict the perceived quality of stippled representations based on the properties of an input image. Our results are related to Weber–Fechner's law from psychophysics and indicate a logarithmic relation between numbers of points and perceived abstraction quality. We give guidance for the number of stipple points that is typically enough to represent an input image well.

CCS CONCEPTS

• **Computing methodologies → Non-photorealistic rendering**; **Perception**;

KEYWORDS

Visual Abstraction, Quantitative Evaluation, User Study, Perception, Stippling, Non-Photorealistic Rendering

ACM Reference format:
Marc Spicker, Franz Hahn, Thomas Lindemeier, Dietmar Saupe, and Oliver Deussen. 2017. Quantifying Visual Abstraction Quality for Stipple Drawings. In *Proceedings of NPAR'17, Los Angeles, CA, USA, July 29-30, 2017,* 10 pages.
DOI: http://dx.doi.org/10.1145/3092919.3092923

1 INTRODUCTION

The process of creating the appearance of shading by carefully placing dots is called stippling. It is a powerful illustration technique frequently found in areas like archeology and biology [Hodges 2003]. The method can be seen as a form of visual abstraction that removes superfluous details for a more efficient visual communication.

A vast amount of research has been dedicated to the automatic creation of such illustrations, reducing the time required by an artist from days to hours or even minutes [Deussen et al. 2000]. Some real-time techniques exist as well [Pastor et al. 2003]. Many of these methods aim to optimize the blue noise properties of point sets, which is commonly argued to be a quality metric for stippling and pointillism. Example-based approaches, e.g. Isenberg et al. [2006], try to combine human input in form of created tonal maps with fast computer-generated 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 paper focuses on the perceptual evaluation of stipple illustrations with regards to their abstraction quality. We investigate how the number of points is related to the perceived quality of an illustration. An answer to this question is crucial, since most stippling algorithms require the user to manually select the number of stipple dots, which is usually a trade-off between quality, computation time and the problem that with too many small stipples the illustration looses its specific look. Our goal is to help users making an educated choice about this number based on our findings.

Since there is no objective function to judge human perception of such drawings, we conducted a user study based on a paired comparison model. For this, a number of input images was normalized using a tonal percentage measure. Users were then presented the input image along with two stipple drawings created from different numbers of points, with the task to judge which one is the better abstract representation. From many such comparisons it is possible to infer an absolute scale using Thurstone's Model [Woods et al. 2010].

The main contributions in this paper are the following:

- We introduce a tonal percentage measure for comparing stipple drawings from different inputs.
- We perform a study that assesses the subjective perceived quality of abstractions.
- We show that the tonal percentage is related to the perceived quality of abstractions by a log-like behavior.
- We present a model to predict the perceived abstraction quality based on the input image.
- We give guidance for deciding the number of points to be used in computer-generated stipple illustrations.

After reviewing related work, we describe the method used to create our stipple drawings and introduce a normalization method. We outline the theory that describes how we reconstruct absolute scale values from paired comparison data. We then discuss our user study, from design and quality control to analysis aspects. Lastly, we draw conclusions from the study and motivate possible future works.

2 RELATED WORK

Stippling Techniques. Many stippling algorithms are based upon Lloyd's relaxation method [Lloyd 1982]: Deussen et al. [2000] analyze the artistic process behind the creation of stipple drawings and propose an interactive editor based on Lloyd's method to create such drawings much faster. Secord [2002] extends this idea by using weighted centroidal Voronoi diagrams that automatically adapt to the underlying density function of the input image. In order to let stipple patterns follow image features, Kim et al. [2008] constrain Lloyd's algorithm via parallel offset lines. A more general energy term is introduced into Lloyd's relaxation by Deussen [2009]. For a certain amount of energy, aesthetically pleasing point configurations appear. Balzer et al. [2009] optimize blue-noise properties of existing point sets, a characteristic that is commonly believed to describe the quality of stippling illustrations. They propose a method similar to Lloyd's method which utilizes the concept of capacity to avoid unwanted hexagonal substructures formed by Lloyd's relaxation.

Example-based methods use hand-drawn stipple drawings and their dots to create a more faithful reproduction: Based on the analysis of neighborhood relationships in an interactive system, Barla et al. [2006] synthesize different styles of stippling and hatching. Kim et al. [2009] propose a technique to analyze, capture, and reproduce the unique artistic stippling style of an artist based on texture synthesis. Stippling is treated as a greyscale process by Martín et al. [2010; 2011], they create illustrations depending on the spatial output size and resolution from examples

Stipple drawings were also created using completely different approaches: Mould [2007] transforms an input image into a weighted graph with image gradients as weights. This graph is traversed with Dijkstra's algorithm and stipples are placed whenever the sum over traversed edges exceeds a given threshold. Li and Mould [2011] focus on retaining image structure by providing a priority-based error diffusion method that gives high priority to extremal values. The technique of Kopf et al. [2006] generates non-periodic point sets that can be used for tiling, allowing viewers to interactively resize the stipple image. De Goes et al. [2012] formulate the calculation of a Capacity-Constrained Voronoi Tesselation (CCVT) as an optimal transport problem, creating point sets with high-quality blue noise and spectral properties.

Results similar to those of stippling can also be produced by halftoning methods. Pang et al. [2008] present an optimization-based technique that preserves structure and tonal similarities by employing a customized objective function. Li and Mould [2010] enhance structures and contrast by an extension to the Floyd-Steinberg error diffusion algorithm [1976] using an adaptive, contrast-aware mask. For a more thorough survey of stippling we refer to the book of Deussen and Isenberg [2013].

Evaluations. Overviews of different evaluation studies in the field of Non-Photorealistic Rendering are presented by Gatzidis et al. [2008] and Isenberg [2013]. Rivotti et al. [2007] discuss how composition principles like unity, balance, center of interest, and emphasis can be used to support the transfer of information in non-photorealistic rendering and increase its visual quality. Although they present two case studies to show the influence of these principles, they do not include any quantitative evaluation. The effect of NPR on 3D shape perception is evaluated by Cole et al. [2009]. Isenberg et al. [2006] present an observational study to examine the assessment of hand-drawn pen-and-ink illustrations of objects compared to computer-generated non-photorealistic renditions. They conclude that not all NPR algorithms are equally successful in creating good scientific illustrations. Maciejewski et al. [2007; 2008] explore these differences using image-processing techniques. They show that the stipple distribution statistics that affect the aesthetics vary among hand-drawn, computer-generated, and natural stipple textures with human created stippling coming closer to those of natural textures. With the goal of more faithfully replicating the traditional stippling process within the digital domain, Martín et al. [2015] focus on the properties of stipple dots, as well as the dimensions of pens and paper types used in artistic practice. Based on a user study, they establish a set of characteristics and conditions for faithfully reproducing traditional stipplings.

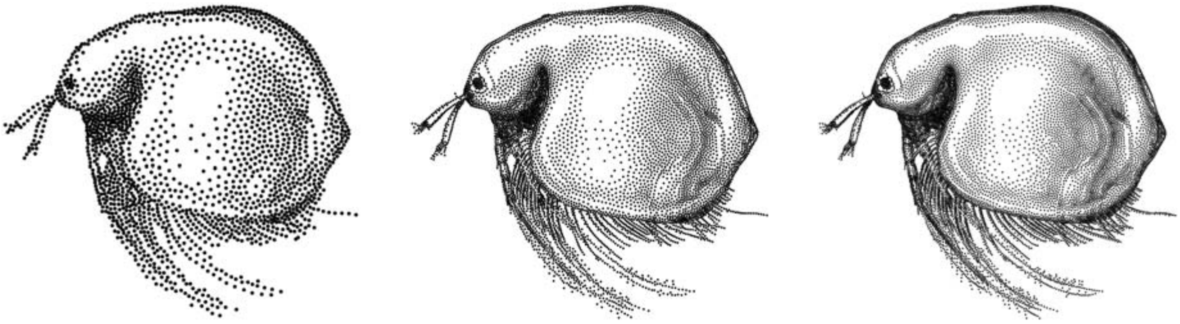

Figure 2: Stippling results created from the same input with increasing tonal percentages (5, 15, 25%), resulting in 1.8k, 5.5k, and 9.2k points (from left to right).

Evaluations on correlation perception in information visualization have been shown to follow Weber's law [Harrison et al. 2014]. Later, it has been argued that a different model is more accurate and general [Kay and Heer 2016].

Many of the discussed works on evaluations focus on the reproduction quality of computer-generated illustrations compared to those created by hand. To the best of our knowledge no work has explored the perceived abstraction quality of these representations depending on the number of dots yet.

3 METHOD

In this section we describe the process behind the creation of the stipple drawings in our study. After presenting the algorithm in a compact form, we introduce our normalization method that relates the number of stipple points to the average tonal value and size of the target image. This makes it possible to compare stipple drawings from different inputs.

3.1 Stippling Algorithm

We use the method proposed by Secord [2002] to generate our stipple drawings. The algorithm makes use of a weighted variant of Lloyd's relaxation method [Lloyd 1982], that 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 most importantly, the number of points can be directly controlled. Other stippling algorithms [Kim et al. 2008, 2010; Li and Mould 2011; Martín et al. 2010, 2011] only offer an indirect control over this parameter.

Following Secord, we create an initial distribution of n points $P = \{P_0, ..., P_n\}$ by rejection sampling. Then, the weighted Voronoi diagram $V = \{V_0, ..., V_n\}$ 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). After the Voronoi diagram is calculated, 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, which we define as the inverted brightness value. This step is repeated for some iterations until the algorithm reaches a state called Centroidal Voronoi Distribution where all points P_i are placed in the weighted centroids m_i of their respective Voronoi cells. The iteration is usually stopped when the average movement of points falls under a user-defined threshold.

3.2 Normalizing Stippling Results

When comparing stipple representations of different inputs we cannot 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 form of normalization would be to normalize all input images with regard to their size and brightness. 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 the fraction of the tonal sum 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 we have an image with 1000 × 1000 pixels and average tonal value of 0.2. Summing up each tonal value results in a tonal sum of 200k. If we choose the number of stipple dots as a tonal percentage of 10% we would end up with 20k points. If all dots would have the size of a pixel, choosing 100% would result in the same number of dots as produced by an error-diffusion algorithm such as Floyd–Steinberg [Floyd and Steinberg 1976]. To faithfully represent the overall tonal value of the input, the stipple size has to be adjusted according to this number of points. The resulting stipple size is then used for all stipples of the current drawing to restrict the number of variables in the user study. Therefore we can divide the overall tonal sum by the number of points and deduce the radius of the stipples from the resulting area.

Stippling images created with increasing tonal percentages are depicted in Figure 2 with 5, 15, and 25% from left to right. In Figure 3 we show three results created for the same percentage from different inputs, resulting in different numbers of points depending on image size and content. Since the church on the right contains darker areas

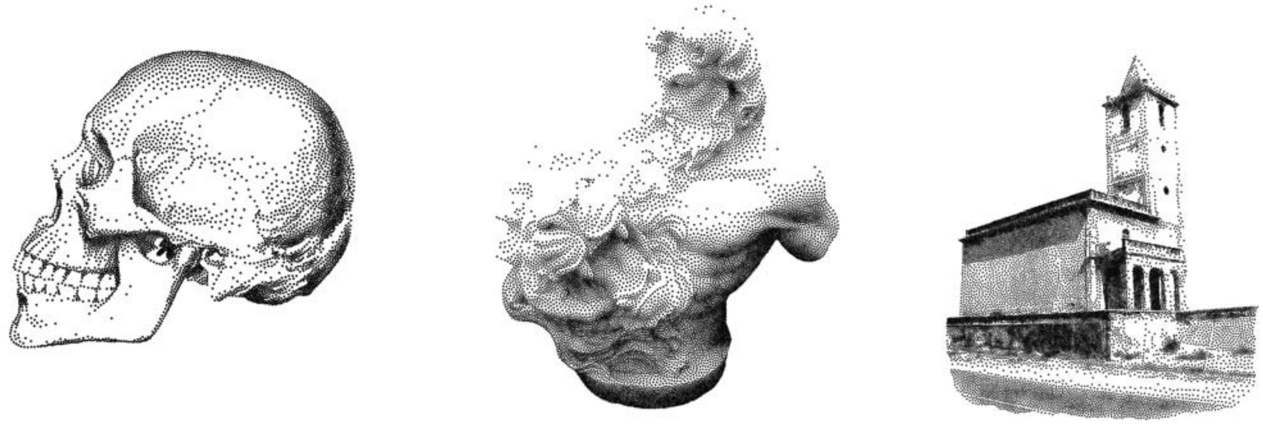

Figure 3: Stippling results created with the same tonal percentage (30%), resulting in 4k, 9k, and 11.2k points (from left to right). The difference in numbers is due to the different size and content of the input images.

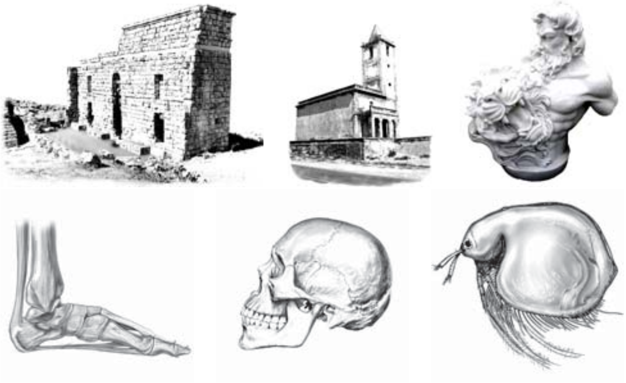

Figure 4: Input images used throughout this paper. Top images from archeology: Acinipo and Church (copyright Domingo Martín Perandrés), Lion; bottom line from biology (taken from [Kim et al. 2009]): Foot, Skull, Water Flea.

Table 1: Input image statistics and their respective values (number of points and radius) with our normalization.

Image	Size	avg. tonal value	5%		10%		15%	
			n	r	n	r	n	r
Acinipo	512×384	0.1903	1871	2.523	3742	1.784	5613	1.457
Church	512×683	0.1605	2807	2.523	5614	1.784	8421	1.457
Foot	512×512	0.0815	1068	2.523	2136	1.784	3204	1.457
Skull	512×512	0.0773	1013	2.523	2026	1.784	3039	1.457
Water Flea	512×512	0.1409	1847	2.523	3694	1.784	5541	1.457
Lion	512×600	0.1474	2264	2.523	4528	1.784	6792	1.457

compared to the other two, it is represented by the highest number of points at the same tonal percentage.

In Table 1 we show some statistics for all input images used throughout this paper, such as their size and average tonal values. In addition, we show their number of points and their respective radius for three different normalization levels. All input image widths were scaled down to 512 pixels in order to have a basis for comparison. The input images can be seen in Figure 4. They were chosen according to the typical application domains where stipple illustrations are frequently used (e.g. in textbooks): top row archeology, bottom row biology.

4 MEASURING ABSTRACTION

Assessing the quality of an image abstraction is a difficult task in general. The same quality score might be interpreted differently by different people. In addition, maximum and minimum scores on a scale may be hard to judge for viewers. Instead, it is much easier to answer which of two given abstractions they consider to be a better representation of a source image. We expect that a higher number of points will always be considered the better representation, but the question we want to address is how this relation changes depending on the number of points. The following subsections describe how we reconstruct abstraction scale values from many such comparisons.

4.1 Thurstone's Model

The result of a paired comparison experiment is a count matrix C that denotes the number of times that each stimulus was preferred over any other stimulus. For n comparisons of stimulus i with stimulus j the count matrix will have the entry $C_{i,j}$, giving the number of times i was preferred over j and $C_{j,i}$, the number of times j was preferred over i, with $C_{i,j} + C_{j,i} = n$. In order to reconstruct absolute score values for the abstraction quality, we employ a variation of Thurstone's law of comparative judgment [1927], a classical method for assigning scale values to stimuli on a one-dimensional continuum from paired comparison data. Thurstonian scaling has widely been used in a variety of areas, ranging from psychology [1967] to subjective preference for video enhancement methods [2010].

According to Thurstone, opinions about a subjective dimension of several stimuli numbered $i = 1, \ldots, m$ are modelled as Gaussian random variables S_i with mean opinions μ_i and variances σ_i^2. This describes the variability in responses to the same stimulus between individuals, as well as when one individual is subjected to the same stimulus repeatedly. Additionally, this accounts for the fact that stimuli of different magnitude can be perceived as equally strong within the subjective dimension to be tested.

When an individual assesses scores of two stimuli relative to each other, it can thus be modeled again as a Gaussian $S_{ij} = S_i - S_j$ with mean $\mu_{ij} = \mu_i - \mu_j$ and variance $\sigma^2_{ij} = \sigma^2_i + \sigma^2_j$ (if S_i and S_j are uncorrelated). In accordance with Thurstone's proposed simplifications, it is common to assume that all $\sigma_i = \sigma_j = \frac{1}{\sqrt{2}}$, so that all $\sigma_{ij} = 1$. Therefore, the probability of a subject to prefer stimulus i over stimulus j is:

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

where $\Phi(\cdot)$ is the standard normal cumulative distribution function (CDF). $P(S_i > S_j)$ can be estimated by the empirical proportion of people preferring S_i over S_j, which can be derived from C as:

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

The mean quality difference μ_{ij} can then be derived from inverting Eq. (1), giving:

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

Here, $\Phi^{-1}(\cdot)$ is 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:

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

For a thorough description of this process and its derivation we refer to the technical report of Tsukida and Gupta [2011]. We augmented Tsukida's implementation to allow non-binary scales and perform the maximum likelihood estimation via a nonlinear programming solver.

4.2 Non-binary scale

In the case of a non-binary scale for judging the quality difference in a paired comparison, the count matrix C has to be augmented to account for the different options. In our study we use a 5-point Likert scale [Likert 1932] indicating strong and weak preference on either side with an additional neutral option. Let S and W be count matrices of the number of times that an option was strongly or weakly preferred, and N be a count matrix of the number of times the neutral option was chosen. Further, let the boundaries between adjacent options on the Likert scale be $\delta = [-\delta_1, -\delta_0, \delta_0, \delta_1]$ and $\mu_{ij} = \mu_i - \mu_j$. The log-likelihood of scale values μ given S, W and N then is:

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

Therefore, the computation of the maximum likelihood solution scale values as given in Eq. (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 a user study on the crowdsourcing platform *CrowdFlower®*[1] to assess the perceived quality of the stippled abstractions. It automatically assigns micro-tasks to workers with different mechanisms for quality control and quality assurance. The micro-tasks consisted of showing one of the input images of Figure 4 together with two of its stippled representations. In the next subsections we will describe how we designed the user study and how we assured the quality of the collected data. The evaluation of the responses will be discussed in the next section.

5.1 Study Design

For each micro-task, crowd workers were shown three images next to each other. As shown in Figure 1, the input was placed in the center with two stippled representations to its left and right. Workers were asked which of the abstractions represent the input image better. For the rating, they were shown a rating scale which is depicted at the bottom of Figure 1. On this scale, a choice for a "much better" preference of one of the two input images is reflected in the S matrix, while a choice for a "slightly better" preference is reflected in the W matrix, and the neutral position counts towards the N count matrix.

In a small pilot study we used pairs of representations with six tonal percentages $\tau_{\text{pilot}} \in \{5, 10, 15, 20, 25, 30\}$. We evaluated a set of all-vs-all comparisons, including symmetric comparisons and comparisons of representations with the same tonal percentage. The results of this preliminary study indicated a logarithmic relationship between τ and the perceived quality of abstraction. Thus, we sampled each input image for the main study accordingly. For the main study, our dataset consisted of six stimuli with 20 representations each, where the tonal percentages were sampled logarithmically:

[1] www.crowdflower.com

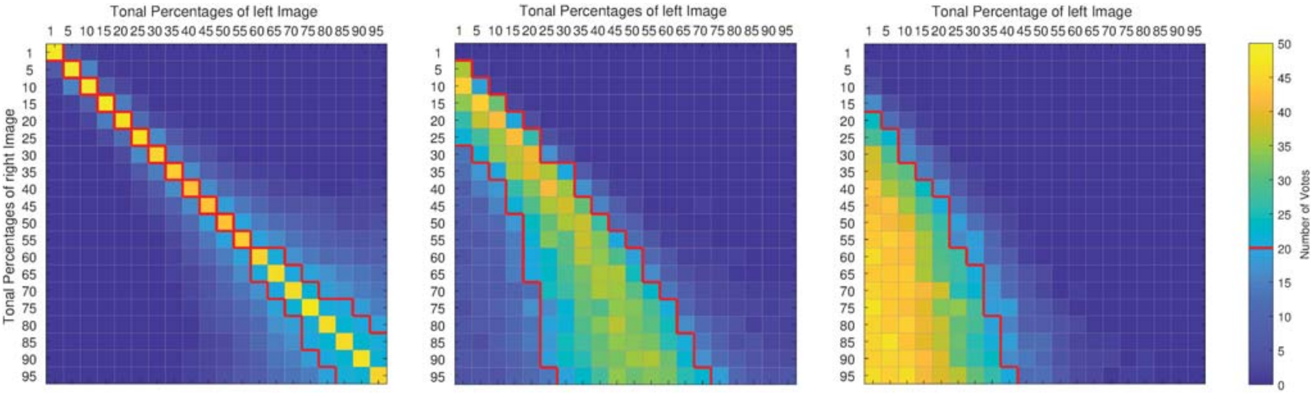

Figure 5: The three count matrices depict vote counts for neutral (N), weak (W) and strong (S) preference options (left to right) on the 5-point Likert scale of all 20 chosen tonal percentages averaged over all 6 stimuli. The red line indicates the region where at least 20 votes were distributed per pair. For the purpose of this visualization we sorted the pairs so that the abstraction with a lower tonal percentage is on the left.

$\tau \in \{1, 2, 3, 5, 7, 10, 13, 17, 21, 26, 31, 37, 43, 50, 57, 64, 73, 81, 90, 100\}$. Since we found no noticeable left-right bias in our analysis of the pilot results, we removed symmetric micro-tasks (e.g. 5% to 10% and 10% to 5%) to reduce the number of overall comparisons. Nonetheless, we randomized the ordering of pairs of stippled representations to avoid any learning effects. Additionally, we ensured that workers were presented with a pair of all six input images before any of them were repeated.

5.2 Quality Assurance and Quality Control

Prior to participation in our experiment, a short introduction into the technique of stippling was given to the workers, including an example illustration. In addition, users were briefly instructed on how they can determine and rate the quality of abstractions. Some factors were listed that influence the judgment of an abstraction such as reproduction of details and matching of shading, as well as the appropriateness of the number or 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 can be considered as a factor when assessing the quality of an abstraction. Additionally, workers had to complete a quiz of ten micro-tasks per stimulus with at least 70% accuracy before being allowed through to the work items.

Workers were allowed to take incremental batches of ten micro-tasks per stimulus. We performed quality control via one hidden test question per stimulus and batch that was a continuation of the quiz. The accuracy on these questions had to remain above 70% throughout the experiment. If users fell below this threshold, all of their answers were discarded. Since ground truth for our questions does not exist, we created two types of test questions based on an appropriate answer distribution principle:

(1) Comparisons of two identical stippled representations were added as test items, where weak preference and neutral options were considered valid answers.
(2) Using the comparisons performed in the pilot study we obtained an estimate for the answer distribution for the respective pairs. We considered answers within the 90% confidence interval of the mean opinion as valid.

Using all of these quality control measures allowed us to detect and exclude 40 potentially fraudulent workers, while 245 workers completed their work accurately. Workers answered an average of 317 answers, with an average test question accuracy of 94.7%.

6 EVALUATION

Our dataset consists of 20 abstractions for each of the six input images with logarithmically spaced integer tonal percentages. We excluded symmetric questions, but included comparisons of a stimulus to itself. In total, this yielded 1260 comparisons for which we accumulated 50 user ratings, under the aforementioned quality control restrictions, resulting in 63000 valid answers.

6.1 Voting Behaviour

Crowd workers had three orders of magnitude (neutral, weak, strong) to choose from, when deciding on the quality differences of any given pair. The vote counts for each of these preferential options (averaged for all stimuli) are displayed in Figure 5. In order to visualize the voting behavior accurately we sorted the pairs so that the abstraction with a lower tonal percentage is on the left.

As expected, the neutral option (left image) was chosen predominantly for those pairs of abstractions that had the same tonal percentages. For higher tonal percentages of both stimuli the discrimination between qualities of abstractions became a harder task, as perceptual differences are less noticeable. Accordingly, an increase in votes for the neutral option can be observed between pairs of high, unequal tonal percentages, as indicated by the increasing width of the region where at least 20 out of the 50 votes were cast, outlined in red. This increase in uncertainty can also be observed in the voting behavior for weak and high preferential options (middle and right image, respectively), where the breadth of the distribution of votes increases similarly.

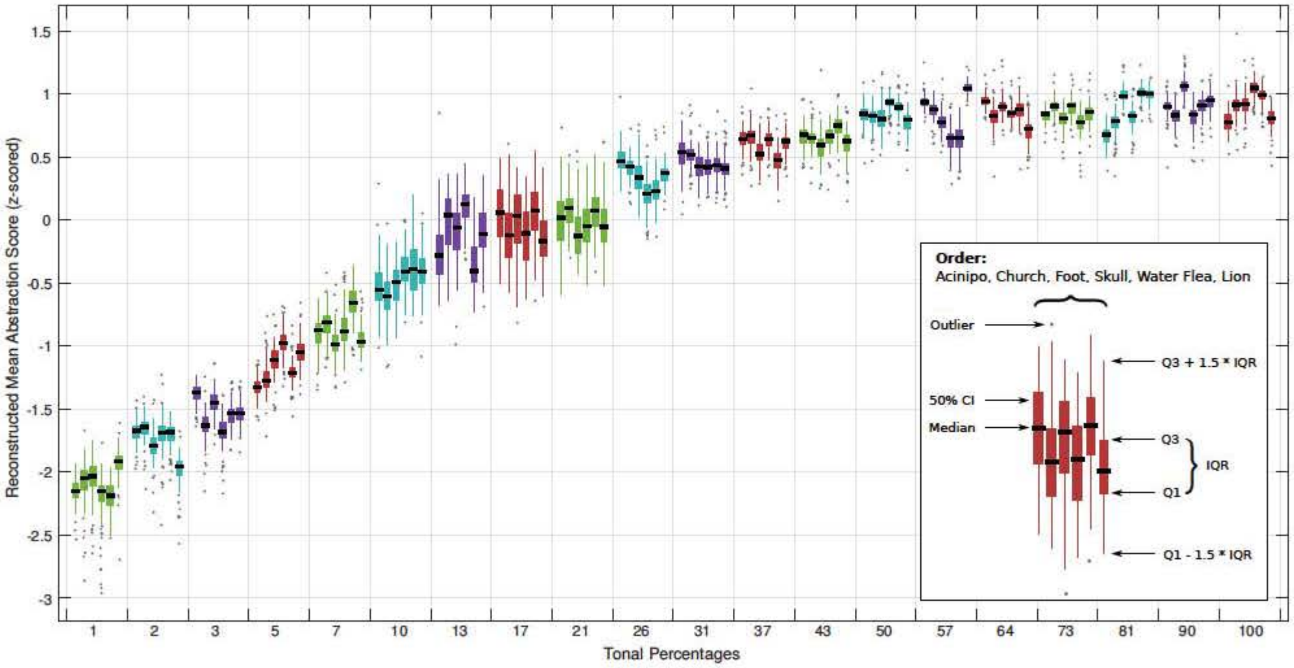

Figure 6: Boxplot for 100 reconstructions of all stimuli grouped by tonal percentages. In this illustration the black bar indicates the median abstraction score, the main body covers the 50% confidence interval, also denoted as IQR. The whiskers extend from the lower and upper quartiles to the last data point in a range of 1.5 · IQR. Outliers are denoted as grey dots.

Table 2: Coefficients and Goodness-of-fit of the logarithmic regression models for all stimuli.

	Stimulus	α	β	γ	R^2	RMSE
1	Acinipo	-2.2197	0.7362	-0.0937	0.9603	0.1943
2	Church	-2.2033	0.7316	-0.0656	0.9560	0.2045
3	Foot	-2.2122	0.7342	-0.0768	0.9597	0.1957
4	Skull	-2.1182	0.7106	0.1588	0.9667	0.1780
5	Water Flea	-2.1581	0.7212	0.0782	0.9700	0.1691
6	Lion	-2.2538	0.7456	-0.1538	0.9689	0.1721
	Mean	-2.1942	0.7299	-0.0255	0.9636	0.1856

6.2 Reconstruction of Abstraction Scores

By applying the reconstruction algorithm proposed in Eq. (3) we are able to compute absolute mean abstraction scores (MAS). Figure 6 shows distributions of 100 such reconstructions for all six stimuli as boxplots. Due to the logarithmic relationship between tonal percentages and MAS that can be observed in this illustration, we propose a logarithmic regression model. The closed form formula to predict the abstraction score from the tonal percentage is

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

where τ is the tonal percentage, and α, β, and γ are content dependent coefficients. The fitted curves alongside the mean abstraction score of the 100 reconstructions is depicted in Figure 7, the coefficients and goodness-of-fit statistics are denoted in Table 2.

The resulting model $\mathrm{MAS}(\tau) = -2.1942 + 0.7299 \cdot \ln(\tau + 0.0255)$ is in accordance with *Weber–Fechner's Law (WFL)* [Fechner 1860] that describes the logarithmic relationship between physical stimuli and human perception. In 1834, German physiologist Ernst Heinrich Weber described that the ratio of the just noticeable threshold of change in stimulus intensity to the intensity of the original stimulus is a constant. Later, Gustav Fechner formulated what is known today as WFL. The differential perception is proportional to the relative change of the stimulus:

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

with I being the intensity of the stimulus, and k is 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 holds for a broad variety of perceptual scenarios, one of which has been proven to be human vision. We refer to Reichl et al. [2010] for an in-depth discussion of the application of the WFL in perceptual domains. New is the fact that this law seems to also apply to the number of drawing objects (stipples) in an abstracted image representation. While already Wertheimer [King and Wertheimer 2004] found out that we see assemblies of many objects as a whole, so far it was not clear that the WFL holds for the perceived quality of abstractions.

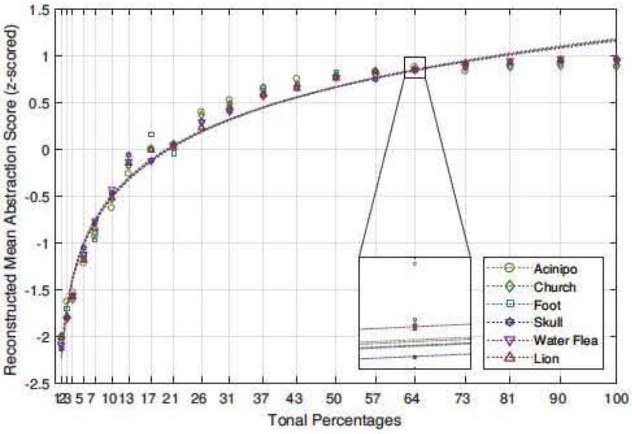

Figure 7: Reconstruction of abstraction scores based on paired comparison data using Thurstonian scaling. The data points are averaged over 100 reconstructions, since the result of two are not necessarily the same, due to random initialization of δ_0 and δ_1. Overlaid are logarithmic curves fitted to the data of all six stimuli, respectively.

6.3 Accuracy of Prediction

In order to evaluate the performance and to verify the generality of our model, we conducted an additional subjective experiment with a different stimulus. It followed the same settings as the main study (cf. Section 5), except that we chose a different set of tonal percentages and the participants were different. We chose 21 tonal percentages, namely $\tau_{\text{eval}} \in \{1, 5, 10, 15, \ldots, 95, 100\}$. Let $\hat{\mu}_{\text{eval}}$ be the z-scored MAS scores obtained through MLE. Then, their distribution has to be fit to the expected distribution imposed by our model. Therefore, we adjust them to have the same population mean and standard deviation as MAS(τ_{eval}). The result of this fitted evaluation dataset is depicted in Figure 8, which shows a good fit of our model to the evaluation data. The model explains 94.32% of the variability of the data, and the RMSE is very low with 0.1844. This indicates that the τ−MAS relationship complies with the WFL.

7 CONCLUSION

We presented a new way of comparing stipple drawings from different inputs by introducing the scale and content invariant tonal percentage measure. Using this measure, we conducted a study that investigates the relation between number of points and abstraction quality of stipple drawings, showing a logarithmic dependency that is 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. We showed that this is also the case for abstract representations, at least for stipple drawings.

Our results help users to choose an adequate number of dots in computer-generated stipple drawings – usually, the number of stipple dots is a user-defined parameter and there is a trade-off between quality and computation time. The reconstructed scale values represent the perceived visual abstraction quality, that is positively correlated with the tonal values. Since our model of the abstraction quality is steadily increasing, 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 weather the increment in number of points, and therefore computation time, is worth the corresponding change in quality. This has to be decided for each application individually. As an example, the quality gain by increasing the tonal value from 10% to 20% is more than three times as high as that from 40% to 50%. If we would have to break down the conclusion to one statement it would be the following: In order to achieve a good perceived visual abstraction quality of your stippled representation, aim for 20-40% of the inputs tonal sum as the number of dots. Less is detrimental to the quality, more only increases it slightly, see the low slope of the model function after this point. In addition to that, choosing a larger percentage also decreases the size of the stipples to a point where single dots become indistinguishable, which is usually not done in stippling.

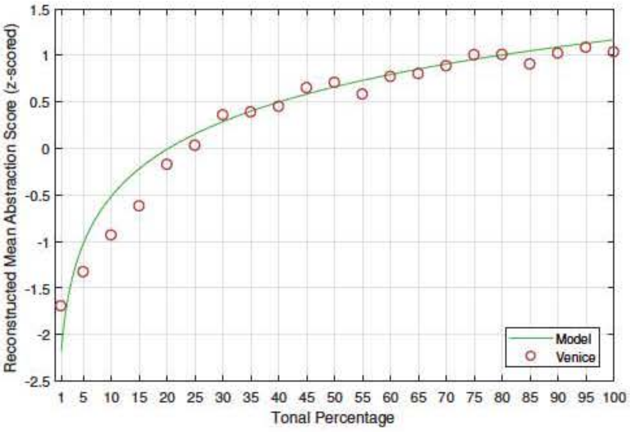

Figure 8: Comparison of abstraction scores of the evaluation dataset plotted in red with the model in green.

There are, however, some restrictions to the applicability of our model. Until now, we do not include semantics into our model: We treat the image with uniform importance and do not vary the stipple size. While a test with three portrait inputs showed that our model did still work well and had a similar RMSE as in our evaluation, we expect this to differ in case of treating certain regions, like the eyes or mouth, with more importance, i.e. distributing more stipples or varying their size. We expect higher perceived quality values for the same number of points in this case. The reasoning behind this is the fact that face-processing is performed by other parts of the visual cortex than the recognition of artifacts or biological objects [Johnson 2005; Kanwisher et al. 1997; Tsao et al. 2006].

8 FUTURE WORK

Future works will include variable point sizes for stippling. This will allow us to use small points in areas with high variance and vice versa, making better use of the overall number of points and creating perceptually adapted distributions. In a next step we will also consider image semantics: for stippled drawings of faces certain regions (eyes, mouth) will be more important with regards to the overall quality than others. Furthermore, we will conduct a similar study to the one presented in this paper for other illustration techniques, like line drawing and stroke-based rendering. There will be a number of additional challenges, since the output complexity of both techniques will be much more difficult to determine: For line drawings it is unclear where one line ends and another begins, and the number of strokes for stroke-based rendering is only meaningful in the absence of overpainting.

ACKNOWLEDGEMENTS

The authors want to thank the German Research Foundation (DFG) for financial support within project A04 and A05 of SFB/Transregio 161.

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. *Computer Graphics Forum* 25, 3 (2006), 663–671. DOI:https://doi.org/10.1111/j.1467-8659.2006.00986.x

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), 9 pages. DOI: https://doi.org/10.1145/1531326.1531334

Clyde H Coombs. 1967. Thurstone's measurement of social values revisited forty years later. *Journal of Personality and Social Psychology* 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), 11 pages. 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*, Oliver Deussen and Peter Hall (Eds.). 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. *Computer Graphics 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 London, London, 45–61. DOI:https://doi.org/10.1007/978-1-4471-4519-6_3

Gustav Fechner. 1860. *Elemente der Psychophysik*.

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 *2008 12th International Conference Information Visualisation*. 475–480. DOI:https://doi.org/10.1109/IV.2008.75

L. Harrison, F. Yang, S. Franconeri, and R. Chang. 2014. Ranking Visualizations of Correlation Using Weber's Law. *IEEE Transactions on Visualization and Computer Graphics* 20, 12 (Dec 2014), 1943–1952. DOI:https://doi.org/10.1109/TVCG.2014.2346979

Elaine R. S. Hodges. 2003. *The Guild Handbook of Scientific Illustration*. Wiley; 2. edition.

Tobias Isenberg. 2013. *Evaluating and Validating Non-photorealistic and Illustrative Rendering*. Springer London, 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, USA, 115–126. DOI: https://doi.org/10.1145/1124728.1124747

Mark H. Johnson. 2005. Subcortical face processing. *Nature Reviews Neuroscience* 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. *Journal of neuroscience* 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 Transactions on Visualization and Computer Graphics* 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 Nineteenth Eurographics Conference on Rendering (EGSR '08)*. Eurographics Association, Aire-la-Ville, Switzerland, Switzerland, 1209–1216. DOI:https://doi.org/10.1111/j.1467-8659.2008.01259.x

SungYe Kim, Insoo Woo, Ross Maciejewski, and David S. Ebert. 2010. Automated Hedcut Illustration Using Isophotes. In *Proceedings of the 10th International Conference on Smart Graphics (SG'10)*. Springer-Verlag, Berlin, Heidelberg, 172–183. http://dl.acm.org/citation.cfm?id=1894345.1894367

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, USA, 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 *ACM SIGGRAPH 2006 Papers (SIGGRAPH '06)*. ACM, New York, NY, USA, 509–518. DOI:https://doi.org/10.1145/1179352.1141916

Hua Li and David Mould. 2010. Contrast-aware Halftoning. *Computer Graphics Forum* 29, 2 (2010), 273–280. DOI:https://doi.org/10.1111/j.1467-8659.2009.01596.x

Hua Li and David Mould. 2011. Structure-preserving Stippling by Priority-based Error Diffusion. In *Proceedings of Graphics Interface 2011 (GI '11)*. Canadian Human-Computer Communications Society, School of Computer Science, University of Waterloo, Waterloo, Ontario, Canada, 127–134. http://dl.acm.org/citation.cfm?id=1992917.1992938

Rensis Likert. 1932. A technique for the measurement of attitudes. *Archives of psychology* (1932).

Stuart Lloyd. 1982. Least squares quantization in PCM. *IEEE transactions on information 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 Third Eurographics Conference on Computational Aesthetics in Graphics, Visualization and Imaging (Computational Aesthetics'07)*. Eurographics Association, Aire-la-Ville, Switzerland, 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 Computer Graphics and Applications* 28, 2 (March 2008), 62–74. DOI: https://doi.org/10.1109/MCG.2008.35

Domingo Martín, Germán Arroyo, M. Victoria Luzón, and Tobias Isenberg. 2010. Example-based Stippling Using a Scale-dependent Grayscale Process. In *Proceedings of the 8th International Symposium on Non-Photorealistic Animation and Rendering (NPAR '10)*. ACM, New York, NY, USA, 51–61. DOI:https://doi.org/10.1145/1809939.1809946

Domingo Martín, Germán Arroyo, M. Victoria Luzón, and Tobias Isenberg. 2011. Scale-dependent and example-based grayscale stippling. *Computers & Graphics* 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, Switzerland, 103–115. http://dl.acm.org/citation.cfm?id=2810002.2810007

David Mould. 2007. Stipple Placement Using Distance in a Weighted Graph. In *Proceedings of the Third Eurographics Conference on Computational Aesthetics in Graphics, Visualization and Imaging (Computational Aesthetics'07)*. Eurographics Association, Aire-la-Ville, Switzerland, Switzerland, 45–52. DOI:https://doi.org/10.2312/COMPAESTH/COMPAESTH07/045-052

Wai-Man Pang, Yingge Qu, Tien-Tsin Wong, Daniel Cohen-Or, and Pheng-Ann Heng. 2008. Structure-aware Halftoning. *ACM Trans. Graph.* 27, 3, Article 89 (Aug. 2008), 8 pages. DOI:https://doi.org/10.1145/1360612.1360688

O. M. Pastor, B. Freudenberg, and T. Strothotte. 2003. Real-time animated stippling. *IEEE Computer Graphics and Applications* 23, 4 (July 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 *2010 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 Third Eurographics Conference on Computational Aesthetics in Graphics, Visualization and Imaging (Computational Aesthetics'07)*. Eurographics Association, Aire-la-Ville, Switzerland, 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, USA, 37–43. DOI:https://doi.org/10.1145/508530.508537

Louis L. Thurstone. 1927. A law of comparative judgment. *Psychological review* 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,