

Tent-pole spatial defect pooling for prediction of subjective quality assessment of streaks and bands in color printing

D. René Rasmussen

Xerox Corporation
800 Phillips Road
M/S 128-27E

Webster, New York 14580

E-mail: rene.rasmussen@xerox.com

Abstract. An algorithm is described for measuring the subjective, visual impact of 1-D defects (streaks and bands) in prints. A general approach to measurements of spatially localized image defects is described, attempting to directly match three stages of processing by the human observer: formation of the perceived image, identification of individual defects, and pooling of the visual defect magnitudes into an overall assessment. The emphasis of the discussion is on the method of pooling of multiple streaks and band defects. It is demonstrated that the commonly used Minkowski pooling method does not satisfy the basic criteria necessary for this application, and a tent-pole pooling method is defined and analyzed. A complete algorithm for measuring streaks and bands, which uses tent-pole pooling, is described. The performance of the algorithm is demonstrated by comparison to results from new and independently collected subjective ratings. © 2010 SPIE and IS&T.
[DOI: 10.1117/1.3280248]

1 Introduction

Streaks and bands remain among the most challenging defects for digital printing technologies, both in terms of the difficulty of eliminating the defects from a design and manufacturing perspective and in terms of their impact on the perceived quality of printed documents. Even high-quality offset printing is not immune to these types of defects. A print with a pattern of bands where CIE (Commission International de L'éclairage) L^* varies sinusoidally with an amplitude as low as 0.2 can be objectionable at certain spatial frequencies, thus placing extremely tight tolerances on the printing process. There is significant literature on determination of perceptibility and acceptable levels of specific streaks and bands defects,^{1,2} including efficient softcopy methods with simulation of printer banding defects.³

The topic of this paper is instrumented, analytical measurements (in short, “measurements”) of streaks and bands, that correlate with human subjective assessment of the print samples. Note that both the measurements and subjective assessments reported in this paper are performed on test charts, where a perfectly produced print sample would display a large area of absolutely uniform color. Although this

kind of assessment is artificial in some ways, it is a very common type of assessment and is very important to printer manufacturers.

The mechanisms that cause streaks and bands are many and varied, and with each mechanism is a specific, sometimes unique, appearance, or defect “look.” Such mechanisms include, for example, weak or missing jets, poor motion quality, and contamination of electrophotographic charging elements. For any given mechanism and defect look, it is relatively easy to develop a measurement that correlates with visual perception of the defect, as the magnitude of the defect varies. For example, if the defect is a sinusoidal variation of L^* with a 10-mm period, a simple measurement of the amplitude will suffice, and may be useful for optimizing the design of components that are responsible for the defect. Such measurements are of limited value, however, since comparisons often must be made between defects with varied looks. A good example of this is when the measurement is to be used for competitive benchmarking or for any other comparison that involves significantly different printing technologies. The International Committee for Information Technology Standards (INCITS) W1.1 standards activity has focused on exactly this subject:⁴ to define measurements of print quality attributes that correlate with visual perception, even when the print samples span across many printing technologies, i.e., for technology-independent measurements.⁵ The INCITS W1.1 macro-uniformity activity addresses print quality defects that include streaks and bands, but also 2-D macro-uniformity defects such as mottle. That activity has not been able to identify existing measurements for overall macro uniformity that are technology-independent and correlate well with human visual perception. The task is a bit easier when only 1-D defects such as streaks and bands are considered, and that is the subject of this paper.

This paper demonstrates general principles related to technology-independent, perceptually correlated measurements of spatially localized image defects, and then exemplifies those principles via a specific measurement algorithm for streaks and bands.⁶ The algorithm attempts to directly represent the subjective evaluation process by three stages: (1) formation of the perceived image, (2) identification and assessment of individual defects, and (3) pooling

Paper 09117SSR received Jul. 8, 2009; revised manuscript received Oct. 15, 2009; accepted for publication Nov. 10, 2009; published online Jan. 21, 2010.

of the visual defect magnitudes into a single numerical overall assessment of the print sample. It is a conjecture of this paper that these three stages describe the overall subjective assessment well. Section 2 examines “defect diversity” of a set of print samples, then goes on to discuss the three major stages of the visual assessment of streaks and bands. Section 2 further discusses how to quantify the individual defects in preparation for pooling of multiple defects, and presents a psychophysical experiment to determine how to modulate defect amplitudes before pooling.

Section 3 examines the suitability of existing pooling models for the streaks and bands application. The effect, colloquially called the tent-pole effect, for judgment of the overall quality of an image with multiple attributes of impairment, is the effect that the observer tends to focus on the worst attribute, such that other attributes are given less significance than if they had been present as the only attribute of impairment. The common use of Minkowski summation of impairment attributes is a testament to the tent-pole effect, since this form of pooling ensures that the largest value receives relatively greater weight. Wang and Shang⁷ investigated several alternative spatial pooling strategies, including a local distortion-weighted pooling, also based on the observation that the worst areas of the image should be given larger weight. They found success with application to assessment of pictorial image defects. However, Sec. 3 will argue that neither Minkowski summation nor the local distortion-weighted pooling work well for the streaks and bands application. Section 3 introduces “tent-pole defect pooling” as a method to model the last of the three stages and explains a psychophysical experiment used to determine a reasonable “tent-pole function.”

Section 4 explains a particular implementation, the visual streaks and bands (VBS) algorithm, of the principles discussed in Secs. 2 and 3. Section 5 explains a subjective evaluation method and presents a comparison between VBS measurements and subjective assessments. The subjective assessments were conducted by the WG4 committee of Japan, in the context of the International Organization for Standardization/International Electrotechnical Commission (ISO/IEC) SC28/WG4 activity toward development of ISO 24790, and the Japanese WG4 committee kindly provided data reported in Sec. 5 of this paper. The comparison covers a large, diverse set of print samples, which were collected and evaluated visually by the WG4 group, entirely independent of the earlier development of the VBS algorithm. Finally, Sec. 6 discusses known limitations of the measurement approach.

As already mentioned, streaks and bands remain very significant issues for printer manufacturers, and as a consequence there is significant literature that analyzes various aspects of the defects. Mizes *et al.*¹ and Cui *et al.*² investigate perceptibility of random streaking and of inkjet banding, respectively. Several authors focus on the physical characterization of the defects,⁸ and as part of that, consider the difficult problem of how to isolate and characterize specific defect types on real prints that contain a multitude of interacting defects, for example, using combinations of wavelet and discrete Fourier transform (DFT) techniques.^{9–11} These efforts also address the ability to predict subjective assessments based on the physical characterization, however, not in situations where a single subjective

evaluation covers multiple types of defects; that is, they are limited to low defect diversity, as described in Sec. 2. While most previous work addresses perception and measurement of streaks and bands on uniform test charts, Bang *et al.* developed a method to assess discrimination of levels of banding in the more realistic and relevant situation where the defects are imposed on pictorial images.¹²

1.1 Terminology of Streaks and Bands

There does not seem to be universally accepted precise definitions of the terms “streaks” and “bands.” Both terms refer to the presence of linear (1-D) patterns in the image, where the color differs from the surrounding color. The color difference could be purely in lightness, as in the case of “dark streaks,” or it could involve chromatic variations, as in the case of “rainbow banding.” Sometimes, but not always, streaks are taken to mean more or less isolated, aperiodic linear defects running in the process direction, for example, caused by a single missing jet. Bands are sometimes, but not always, used to refer to linear defects that run in the cross-process direction, repeating periodically (perhaps sinusoidally) in the process direction, for example, caused by motion quality problems. The measurement algorithm described in this paper does not make any distinction between isolated or repetitive linear defects, and therefore precise definitions of streaks and bands are not necessary; however, for the verbal discussion it is useful to use both terms largely following the usage described above, but with some differences, as noted here.


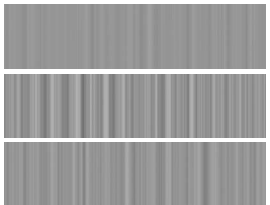
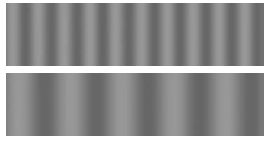
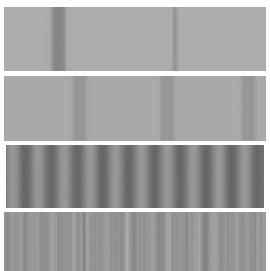
1. *Bands*: The presence of multiple linear defects running either horizontally or vertically in the image (a distinction is not made between process and cross-process directions). The multiple defects would typically follow a periodic pattern in the direction perpendicular to the defects, e.g., a sinusoidal lightness variation. A pattern of such bands is not considered as a single defect, but considered to consist of multiple unrelated defects. An example of bands is given in Table 1 defect class III.
2. *Streak*: A single linear defect running either horizontally or vertically in the image. Examples of streaks are given in Table 1 defect class I (single streak) and class II (multiple streaks).

The VBS algorithm operates on the hypothesis that the overall subjective quality can be modeled by characterizing the image in terms of individual, unrelated linear defects, even in the case of periodic bands; therefore, it is sometimes useful in this paper to refer to the individual linear defects in a pattern of bands, as a “streak.” Note that linear defects that do not run horizontally or vertically are not considered as streaks or bands, and are not within the scope of this paper.

2 Defect Diversity and Factors of the Perceptual Assessment

To construct a successful, perceptually correlated measurement of streaks and bands, it is important to consider carefully the process of visual assessment, even—perhaps especially—those parts of the process that are less well understood, such as those strongly influenced by subjectiv-

Table 1 Four classes of defect diversity.

Class	Example of sample set	Characteristics	Requirements for adequate measurement
I		$g(x) = Af(x)$ Samples have (almost) identical spatial patterns, that vary only by simple scaling of the contrast. One "look."	Any measurement that captures the amplitude of the defect (contrast). Correlation to perceptual assessment is almost certain. A human visual system model is not required.
II		Both the contrast and the spatial pattern vary between samples. However, the samples have essentially identical amplitude spectra (except for an amplitude scaling factor) and appear to have only one "look", with variation only in contrast. Example: Random streaking with $1/f$ amplitude spectrum.	Same as for class I. See text for assumptions.
III		$g(x) = Af(x / \lambda)$ The samples vary both in contrast and in spatial appearance, but the variation of the spatial pattern is limited to changes of the amplitude and of the spatial scale, such that all samples have essentially the same "look."	A human visual system model must be applied, e.g. a simple CSF, but there is no need for complex defect pooling.
IV		Most diverse sample set: - Contrast and spatial scale varies. - Multiple different "looks."	Must encompass both a human visual system model, such as a CSF, and a model of more complex defect pooling.

Note: The classes refer not to a single print sample, but to a set of print samples used to test correlation between measurements and subjective assessments. Each class has different requirements as to which parts of the subjective evaluation process must be taken into account. The print sample set used for the subjective ratings in Sec. 5 belongs to class IV. The functions $g(x)$ and $f(x)$ represent the 1-D luminance variation across two different samples belonging to the same class.

ity. This section and the next discuss general principles, then Sec. 4 discusses how these principles are applied for the VBS algorithm for streaks and bands.

The measurement algorithm described here was developed with the objective to replace (when combined with other algorithms) a time-costly visual assessment procedure called "DAC Macro-uniformity."¹³ DAC Macro-uniformity is defined nearly identically to the macro-uniformity attribute as defined by the INCITS W1.1 macro-uniformity team,⁴ and covers in addition to streaks and bands 2-D defects such as mottle. The performance of the VBS algorithm, when combined with measurements of 2-D defects, in terms of prediction of DAC ratings, has been reported earlier.⁶ In this paper, the focus is strictly on 1-D defects, and the measurement algorithm is tested against visual assessments performed under the ISO 24790 activity. All

three visual assessment procedures are very similar and use the following general procedure. The observer is presented with print samples of a test chart that contains a large nominally uniform region (dimensions vary in the range from 170×170 to 200×300 mm) and is directed to assess the overall appearance of the uniformity of each print sample. The samples typically contain multiple defects in both horizontal and vertical directions, and the observer must assign a single numerical value that represents the overall assessment. More details on the subjective method are given in Sec. 5.

Theories of visual perception operate with several stages of visual processing,^{14,15} starting with the physical optics of light forming an image on the retina, followed by increasingly complex stages of image and data processing. After

the retinal image is formed, image-based processing follows, then surface-based, object-based, and category-based processing. The later stages of the visual processing are related to interpretation and mapping to the 3-D physical world, and are less important to the task of streaks and bands assessment when using test charts. For the purpose of this discussion the assessment process is organized into three stages.

Stage 1: the perceived image. The optical imaging on the retina and the initial image-based processing leads to the 2-D perceived image. This stage starts with the physical imaging of the object through the lens of the eye, forming an image with finite resolution on the retina. For a gray-scale image, the result of further image-based processing can to first order be characterized by human visual contrast sensitivity functions.¹⁶ This first stage accounts for effects that are not under our conscious control, for example, Mach bands, which appear as a result of the bandpass nature of the optical and image-based processing. The output from this stage will subsequently be referred to as the perceived image. A relatively high degree of consistency between observers can be expected at this stage, with variations being driven by physiological differences between observers.

Stage 2: defect identification. The analysis of the perceived image leads to conscious decisions about the presence of one or more defects in the image.* In the case of streaks and bands on a test chart, this means defects that observers can point to and verbally describe. For some images, the processing and analysis that take place during this stage may be straightforward, for example, in case of an image that is perfectly uniform except for the presence of a single streak. In other cases, there is not a unique solution to the decomposition of the perceived image into a set of discrete defects, and the observer may perform some degree of interpretation. Human observers, especially expert observers as used for the data in Sec. 5, perform this task very well, relying partly on background knowledge about the types of defects that are commonly encountered. Therefore, a relatively high degree of consistency among observers can be expected even at this stage (at least compared to the far more subjective process in stage 3).

Stage 3: defect integration (or “pooling”). The conscious pondering of how the collection of defects “adds up” to an overall subjective judgment. The reduction from numerous perceived defects to a single numerical rating involves nontrivial comparisons and highly subjective judgments—processes that presumably take place well outside of the brain’s visual cortex. It can be argued (see Sec. 6), that a meaningful defect integration cannot be performed based purely on the information present in the perceived image, and in the absence of information on how observers perform this task, it could be tempting to ignore

* It is important here to recall that the assessment is performed on nominally uniform test targets, such that the observer’s attention is not distracted by image content, but is solely focused on the defects. It is under that condition only, that we conjecture that observers will tend to identify specific defects.

the complexity of this stage, and model it in a simplistic manner. However, meaningful or not, the defect integration is performed, and a reasonable model for this stage is critical to an overall well-correlated measurement.

In summary, consistent with these three stages the measurement should be composed of the following steps:

1. digitization of the hardcopy and conversion of the digital image values to a colorimetric space
2. modulation of the image to obtain a representation of the perceived image
3. identification of individual defects from the perceived image, and characterization of each defect in a manner that lends itself for subsequent pooling (e.g., numerical characterization of amplitude, size etc.)
4. defect integration (pooling) to obtain a single numerical rating of the sample

This is the general recipe for measurements of spatially localized defects that is used by the VBS algorithm described in Sec. 4.

2.1 Classes of Defect Diversity and Complexity of Assessment

One of the most successful applications of the human contrast sensitivity function (CSF) to print quality measurements is that of graininess,^{17–19} obtained by integration of the CSF-weighted Wiener spectrum. For streaks and bands, this approach does not work, because it ignores the phase relationships, and therefore ignores stages 2 and 3. In the case of graininess, even though the observer can in fact perceive individual defects, in the form of grains, that is not how the typical subjective judgment of graininess is performed—rather, the observer judges graininess as a single entity that is spatially distributed and homogenous at large scales. Therefore, for assessment of pure graininess, stage 3 is trivial. As mentioned previously, the literature covers methods that produce good correlation between measurements and perception of streaks or bands, but often demonstrated only for specific applications using a set of prints with limited or unknown diversity of defects. Defect diversity refers to the amount of variation of the “looks” of the defects that are present in a given set of prints that are used to test correlation between measurements and observations. Table 1 illustrates four significantly different classes of defect diversity. For streaks and bands, there are important examples of each class.

As an example consider random streaking. Random streaking in printers is often characterized by a $1/f$ noise spectrum,¹ within a certain range of spatial frequencies. Assuming that the image size is large enough compared to the lowest spatial frequency, and assuming that the overall contrast is large enough that a very large number of streaks are above the perception threshold, then different images of the same normalized amplitude spectrum will to the observer appear to have the same “look” and will appear to differ significantly only in terms of the overall contrast (see Table 1, class II). Notice, however, that at a detailed level, the images would differ significantly. Given that the samples are characterized by essentially identical amplitude spectra (except for an amplitude scaling factor), a good correlation to visual assessments can be obtained solely

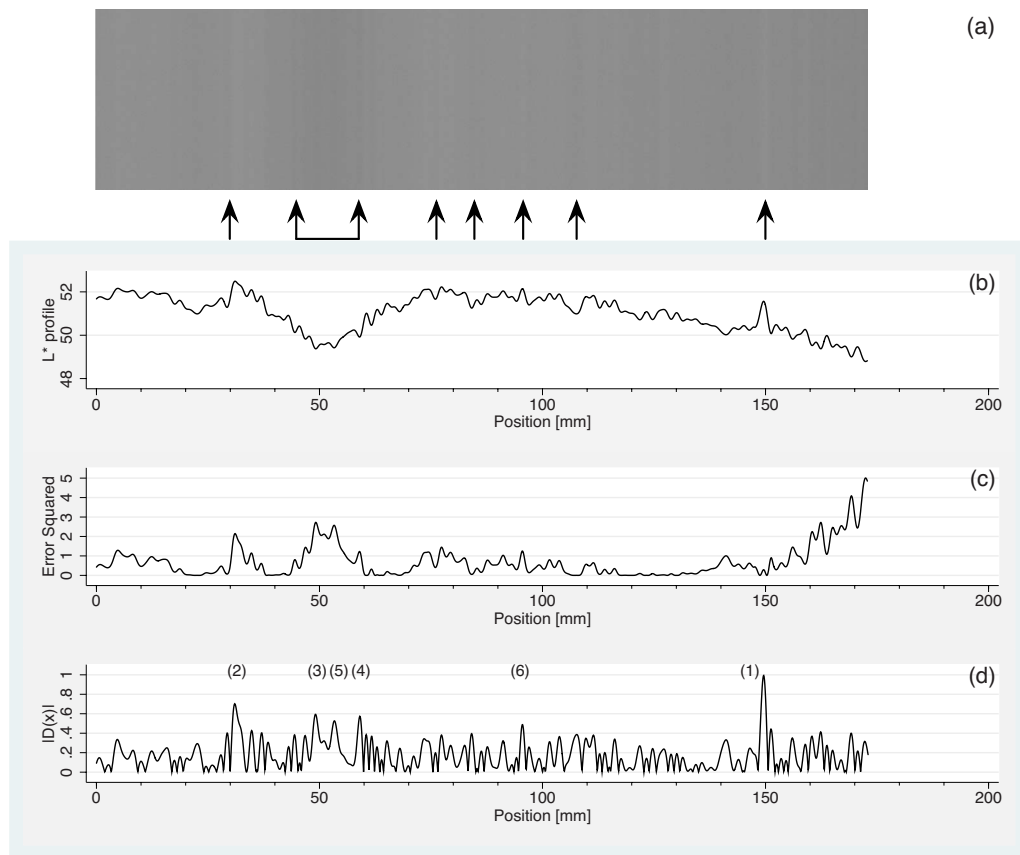


Fig. 1 (a) Segment of an image displaying vertical streak defects. (b) The L^* profile of (a) in the horizontal direction. The single most prominent defect is the narrow light streak near position 150. That defect and a few of the other visually prominent defects are marked by the arrows. (c) The pointwise error squared, calculated by squaring the deviation from the mean L^* . (d) The absolute value of the deviations from the mean of the “perceived profile” obtained by bandpass filtering the profile in (b). The order of the defect magnitudes, as used for tent-pole integration, is indicated (the separation into three frequency bands, explained in Sec. 4, is not illustrated here).

from a measurement of the overall contrast, even without taking the CSF into account. Such a sample set belongs in class II. If the amplitude spectra are allowed to vary from sample to sample away from the $1/f$ distribution, then it will be necessary to take the CSF into account, and this situation—similar to graininess—belongs in class III.

During development of a printer, when technology parameters are varied for optimization, the defect diversity observed in the print samples often belongs to classes I, II, or III. When limited to such cases, it may be relatively easy to find a measurement that correlates with visual assessments. The challenge comes, as noted earlier, when faced with class IV defect diversity, which is the case for competitive benchmarking and for technology-independent measurements. For class IV defect diversity, it is critical to directly address defect detection and pooling. Section 2.2 discuss an experiment and a method to quantify individual defects, and section 3 will discuss pooling of the defects.

2.2 Quantification of Individual Defects in the Perceived Image—The Visual Defect Magnitude

The remainder of this paper until the subjective assessment in Sec. 5 treats only the case of gray-scale images, rather

than full color images. The concepts can be readily extended to color images. Print samples can be digitized, e.g., with a flatbed scanner, and the image converted into CIE L^* lightness space. However, due to the bandpass luminance CSF, there will generally be significant discrepancies between this “measured image” and what the human observer sees, the “perceived image.” Figure 1 illustrates the criticality of taking the CSF into account, since direct measures of the magnitude of deviations from the mean of the L^* profile do not correlate well with perceived defects.

The CSF depends on many factors of the sample and viewing conditions.¹⁶ The VBS algorithm is designed to use a slightly modified interpretation of the CSF, which in the remainder of this paper will be called the quality impairment function (QIF). The rationale for this is as follows. The threshold CSF is a measure of the thresholds [e.g., 50% just noticeable difference (JND)] of perceptibility of sinusoidal luminance variations—thus, by definition is not necessarily applicable to suprathreshold luminance variations. Even contrast-matching sensitivity functions, determined in the suprathreshold regime, cannot be directly interpreted to assign a measure of quality impairment to an individual streak defect, because the question of matching contrast is fundamentally different than the question of

matching quality. Consider two printed pages, each with a single streak in the shape of a full cosine period (that is, the edges are soft) on an otherwise perfectly uniform page. On one page the streak is 2 mm wide, on the other it is 100 mm wide. For the VBS algorithm, the key question is at which relative amplitudes observers would rate the two pages to have the same overall quality. Even if the CSF-modulated amplitudes of the streaks match, that does not imply that the overall quality of the page would be rated the same, because the narrow streaks impact only a very small part of the image.

2.2.1 Experiment to determine QIF

A survey was conducted to directly address this question, with the goal of being able to assign a visual defect magnitude δ to each individual streaklike defect. Prints with simulated dark streak defects were generated, using a photoreality ink jet printer. The prints were letter sized with a uniform base gray with $L^*=L_B=75$. The prints were presented to observers using a mask, such that only a rectangular subregion of the print was visible. The width (x direction) of the visible region was held constant at 170 mm. Several values in the range from 20 to 170 mm were used for the height E of the visible region, thus varying the visible length of the streak. The image content was given as follows:

$$L^*(x, y) = L_B - \frac{A}{2} \{1 + \cos[2\pi(x - x_0)/w]\} \quad (1)$$

for $x \in [x_0 - w/2, x_0 + w/2]$,

and $L^*(x, y) = L_B$ elsewhere, resembling the images in Table 1, class I. Here x_0 is the center of the streak and was varied pseudorandomly from sample to sample, but in such a way that there was always at least 10 mm of uniform region visible on each side of the streak. Also, A is the amplitude from the base L_B to the L^* minimum at the center of the streak, and w is the full width of the streak. The survey used eight different streak widths: $w = 1, 2, 4, 8, 16, 32, 64,$ and 125 mm. The amplitudes were adjusted depending on the streak width, to range from barely perceptible, to clearly perceptible (several L^* units).

Observers were directed to rate the quality of each sample, imagining that it would be used for a document cover page that was intended to be uniform gray. Two anchor samples were displayed corresponding to the worst quality level with anchor value 1, and the highest quality level (no perceptible streak) with anchor value 10. The samples were viewed at approximately a 40-cm viewing distance, in a well-lit office, but otherwise uncontrolled viewing environment. The samples were rated by 20 observers, and the median of the 20 individual ratings was used as the measure of the sample quality.

The QIF was then determined as follows, in a manner that is consistent with its application to the VBS algorithm, including pooling of multiple defects. Each print sample was scanned with a calibrated drum scanner, and the lightness profile $L(x)$ across the streak, was calculated (this step is performed to eliminate the effect of discrepancies between the intended and actual image content, caused by

Table 2 Parameter values used in Eq. (2) for QIFs optimized for 20 and 170 mm heights, respectively (these functions are shown in Fig. 3).

	Optimized for	a	b	k	f_0
QIF 1	$E=20$ mm	0.553	0.40	1.80	0.074
QIF 2	$E=170$ mm	0.617	0.40	1.33	0.074

limitations of the ink jet printer). Assuming a QIF of the following form, where f is the spatial frequency in cycles per millimeter (c/mm):

$$\text{QIF}(f) = a + b \tan^{-1}\{k[\log(f/f_0)]\} \quad \text{for } f < 1 \text{ c/mm.} \quad (2)$$

A filtered profile was calculated as $h(x) = \mathcal{F}^{-1}\{\text{QIF } \mathcal{F}[L(x)]\}$, where \mathcal{F} represents the Fourier transform. The VBS algorithm is designed to exclude high-frequency defects (see Sec. 4), so the domain $f > 1$ c/mm was not used. The visual defect magnitude of the streak δ was then defined as the positive difference from the minimum value to the base level of the filtered profile, that is, $\delta = h(0) - \min h(x)$. The goal was to find a QIF such that δ is proportional to the subjective judgment of quality impairment, independently of the spatial shape of the streak. This means, in particular, that when δ is plotted against the quality rating, it should form a single curve, independent of the streak width. In general, there is no guarantee that a QIF can be determined such that this is the case, but the parameters of Eq. (2) can be optimized to provide a linear fit with the smallest mean squared error.

This parameter optimization was performed for both $E = 20$ mm and $E = 170$ mm, leading to two different QIFs, as shown in Table 2. Figures 2(a) and 2(b) show the visual defect magnitudes δ calculated with these two QIFs, both plotted versus the subjective quality ratings for $E = 20$ mm. Figure 2(a) shows, as expected, that when the QIF is not optimized for the correct height E , then the data points corresponding to different streak widths do not form a single curve. Figure 2(b) shows the width-independent function obtained when the parameters for the QIF are optimized for $E = 20$ mm (QIF 1 in Table 2). Figure 2 also illustrates the importance of taking the height E into account when performing subjective experiments for streaks and bands, since small values of E would significantly underestimate the impact of wide, relative to narrow, streaks when viewed on a full page.

In a similar manner, an optimal QIF was determined for the case the $E = 170$ mm mask. The optimal QIFs for the two cases are graphed in Fig. 3. Note that due to the streak widths used in the experiment, the QIFs were determined only for frequencies less than 1 c/mm, and were arbitrarily normalized to ≈ 1.0 at $f = 1$ c/mm.

The results described in this section provide the basis for assigning an individual streak defect with a scalar value δ called the visual defect magnitude, in such a way that δ is proportional to the perceived impairment of the overall quality. The next section discusses general aspects of how to pool the effects of multiple defects that are present on a single page. The details of how multiple defects are identi-

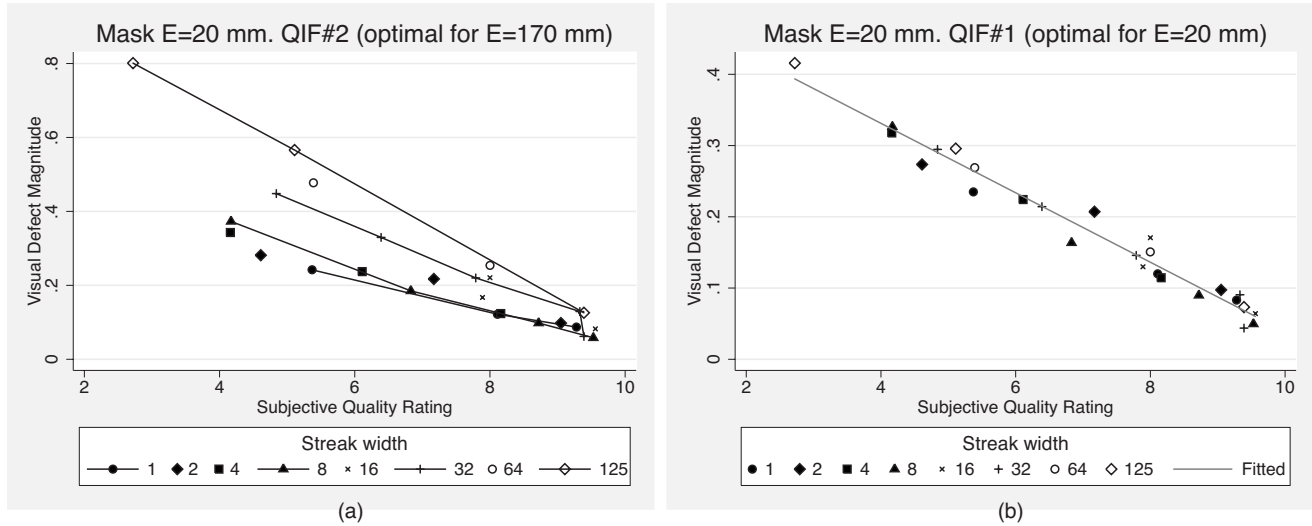


Fig. 2 Calculated visual defect magnitudes versus subjective quality ratings for mask $E=20$ mm, for two different QIFs: (a) defect magnitudes calculated with the nonoptimal QIF 2 (optimized for $E=170$ mm), leading to width-dependent curves; and (b) defect magnitudes calculated with a QIF optimized for $E=20$ mm, leading to a width-independent relationship.

fied and characterized from L^* profiles, are postponed until the discussion of the VBS algorithm in Sec. 4.

3 Tent-Pole Pooling of Discrete Defects

Going back to the perceived image, there are several seemingly reasonable ways to proceed to calculate measures that may correlate with an overall subjective rating. There is significant literature on advanced models of the visual system, which can successfully be used to predict perception of defects in pictorial images, useful, for example, for evaluation of display systems and compression algorithms.²⁰⁻²² The models take into account advanced aspects of the human visual system, beyond the CSF, and can calculate an image detection map, where each pixel value represents the probability that a difference between the original and the processed image will be detected at that location in the image. Norms of the image detection map

(e.g., max-norm, L_p -norms) can then provide measures of the overall image degradation. In the 1-D case of streaks and bands, the “perceived profile” $h(x)$ can be used in a similar manner. We can calculate the perceived profile as $h(x) = \mathcal{F}^{-1}\{\text{CSF } \mathcal{F}[L(x)]\}$. The deviation from the mean is calculated as

$$D(x) = h(x) - \bar{h}, \quad (3)$$

where \bar{h} is the mean of $h(x)$. The absolute value of $|D(x)|$ is illustrated in Fig. 1(d). According to the simple CSF model of the visual system, a location in the image where $|D(x)|$ is relatively large, is perceived as more different from the mean; that is, it contributes significantly to the perception of overall nonuniformity. An overall measure could then be obtained by integration of the profile over the length L :

$$\|D\|_p = \left(\int_0^L |D(x)|^p dx \right)^{1/p}, \quad (4)$$

where the exponent p controls to what extent larger deviations dominate compared to smaller deviations. Variations on this theme exist, for example, introduction of a perceptibility threshold, such that the integration is effectively limited to regions where $D(x)$ exceeds that threshold. However, as discussed in the previous section, there is nothing in the theoretical or experimental foundation for the CSF, or the construction of $D(x)$, that suggests that Eq. (4) would yield a particularly good measure of the overall quality. On the contrary, there are clear problems with such an integration, as illustrated by Fig. 1(d). For $p=1$, the integral is the area under the curve, which means that the width of the defect, as represented by $D(x)$, has a significant impact [it was precisely for this reason that the previous section introduced the QIF to allow a more direct interpretation of $D(x)$].

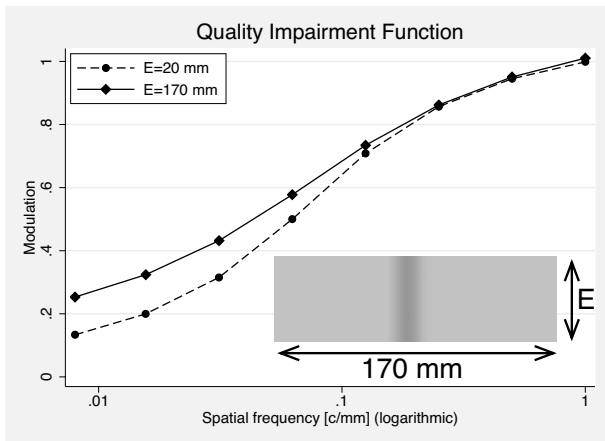


Fig. 3 QIFs tuned for $E=20$ mm and $E=170$ mm, respectively. The VBS algorithm eliminates high-frequency signals, so the QIFs were not determined at higher frequencies.

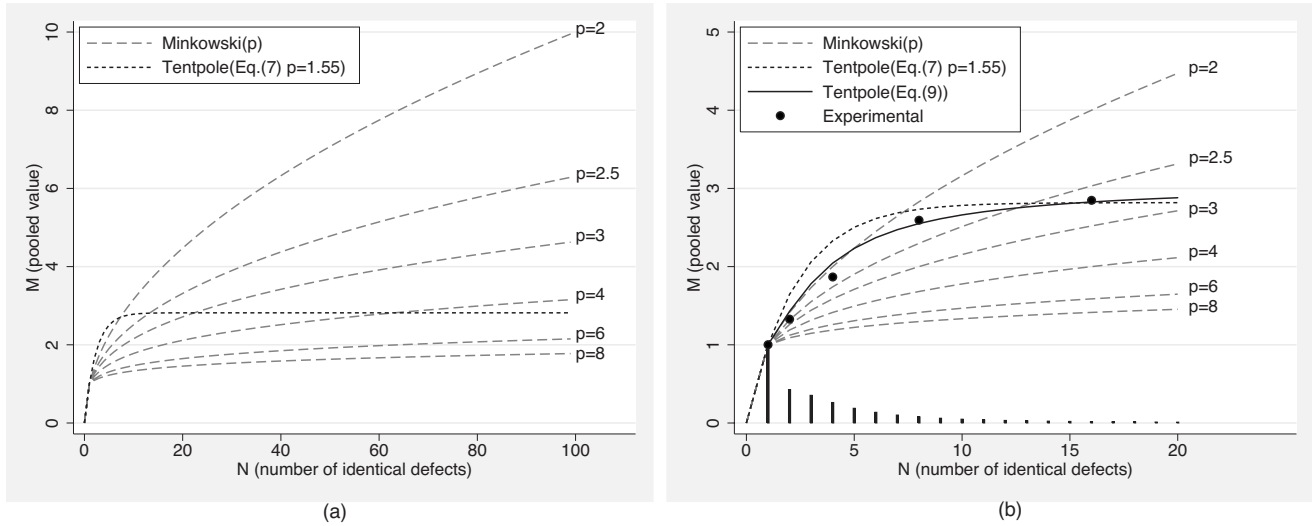


Fig. 4 Pooling of N identical defects each of magnitude 1. The first axis indicates the number of defects, while the second axis is the measure of the overall impairment (the pooled value). (a) Minkowski summation and tent-pole pooling (see Sec. 3.2) over a potentially large number of defects and (b) the range up to $N=20$ including experimental data as described in the text. The solid line is the cumulative tent-pole function for Eq. (9), while the vertical bars show the corresponding tent-pole function $T(N)$.

3.1 Requirements to Pooling Methods

To proceed, we consider how visual defect magnitudes may be combined in cases where a single print contains more than one defect. Three assumptions are made about the overall subjective assessment process:

1. (A1) Assume that the observer identifies a finite set of N discrete defects in the perceived image, and that each defect is associated with a certain magnitude δ_i according to how the defect would have been assessed, had it appeared in isolation on an otherwise perfect image. The δ_i values are interpreted as the visual defect magnitudes described in Sec. 2.2.
2. (A2) Assume that the subjective assessment of overall quality Q does not depend on the spatial arrangement of the individual defects. This assumption is at least somewhat reasonable, since visual spatial processing was taken into account during stage I, although masking of one defect by another has not been taken into account. The perceived defects in the image are then completely characterized by a finite series $\Omega = \{\delta_i\}_{i=1}^N$.
3. (A3) Finally, assume that the subjective assessment of overall quality is a function only of the series $\Omega = \{\delta_i\}_{i=1}^N$, so that the overall quality impairment is $\Delta Q = \Delta Q(\Omega)$.

The equivalent of Eq.(4) for a series of discrete defects is the commonly used Minkowski summation:

$$M_p = \left(\sum_{i=1}^N \delta_i^p \right)^{1/p}, \quad (5)$$

which has been reported to provide a good model for the combined effect of sub- or near-threshold stimuli,²³ as well as for multiple suprathreshold stimuli,²⁴ and often is used

with exponents p in the range from 2 to 4. The remainder of this section argues that Minkowski summation is not suitable for the streaks and bands application, and proposes an alternative based on experimental data.

In the limit of large p , M_p approaches the max norm. It is well known that subjective assessments of overall quality tend to be dominated by the worst defect, and the max norm is the extreme representation of that effect. However, it is apparent that if the magnitude of any one defect is increased (even one that is smaller than the largest defect), then the overall quality must decrease, if even by a very small amount. In particular, an image with two identical defects must have worse quality than the image with just one such defect. Minkowski summation with a sufficiently large p value will suffer from the same problems, but smaller values of p provide defect pooling that is acceptable from this perspective alone. It can also be expected that the overall impact of two defects is less than the sum of the defect magnitudes. The Minkowski summation represents this well, as illustrated in Fig. 4(a). This graph shows M_p of Eq. (5) as a function of the number of defects, in the case where all defects have identical magnitude $\delta_i=1$ (where M_p reduces to $N^{1/p}$). For the application to streaks and bands it is important to consider relatively large numbers of defects. For example, a page that is 200 mm wide with 5-mm sinusoidal banding, contains 40 periods, which can be interpreted as 40 light streaks and 40 dark streaks, that is, a total of 80 defects. Another example is given by Fig. 1(d), which displays ≈ 90 local maxima, each interpreted as an individual defect. These examples support the following postulate, that the defect pooling must asymptote to a constant value rather quickly as the number of defects (even with identical magnitude) increases. If an image contains numerous but very faint streaks, there is a limit to how poor the sample will appear, regardless of the number of such faint streaks; it will always appear better than a

sample with a single severe streak. If it is required that the pooled value of N defects of magnitude δ increases only insignificantly ($\ll \delta$) as N increases from, say, 50 to 100, then Fig. 4(a) shows that Minkowski summation with $p \leq 3$ is disqualified.

Let $\Delta Q(N)$ denote the subjective impairment of overall quality due to N defects of identical visual magnitude $\delta = 1$. We then define the tent-pole function, $T(N)$, as the incremental impairment from $N-1$ to N defects; that is, $T(N) = \Delta Q(N) - \Delta Q(N-1)$, where $N \geq 1$. By definition $T(1) = 1$. The curves in Fig. 4 represent the cumulative tent-pole function $C_T(N) = \sum_{i=1}^N T(i)$, which correspond to the overall impairment $\Delta Q(N)$.

3.1.1 Experiment to determine pooling of defects of identical amplitude

An experiment was conducted to determine the pooling behavior for images with fewer than 20 defects of identical amplitudes. The experiment used a calibrated monitor to display a pair of images of a uniform gray background with one or more vertical streaks. The shapes of the streaks were given by Eq. (1) and the full streak width was constant at 4 mm. The locations of the streaks were pseudorandom, but arranged such that there was at least 10 mm of uniform background between streaks. Thus, the observer was presented with two images, where one image contained N_1 streaks of amplitude A_1 , and the other image contained N_2 streaks of amplitude A_2 . The experiment included comparisons of images with streak count $N = 1, 2, 4, 8, \text{ and } 16$, but the analysis used only the data where one of the two images contained a single streak. The observer was asked to imagine that each image represented the cover page of a document, intended to be uniformly gray, and was forced to select which of the images would produce the best overall quality. Ten observers participated in the experiment, and the results were averaged over the observers.

If $Q(N, A_N)$ denotes the quality of an image with N streaks each with amplitude A_N , then it is possible to determine (by interpolation) the amplitude A_1 required for an image with a single streak to yield the same overall quality: $Q(1, A_1) = Q(N, A_N)$. With A_1 and A_N defined in this way, the ratio of amplitudes can be calculated, as an estimate of the cumulative tent-pole function:

$$C_T(N) = \frac{A_1}{A_N}. \tag{6}$$

The experimental values of $C_T(N)$ are shown on Fig. 4(b) for comparison with Minkowski summation. Minkowski summation with exponent $p > 3$ does not match these experimental data, which combined with the prior observation for $p \leq 3$ rules out Minkowski summation as a viable option for the streaks and bands application.

3.2 Tent-Pole Pooling

As mentioned in the introduction, the common use of Minkowski summation (with $p > 1$) of impairment attributes is a testament to the tent-pole effect, since this form of pooling ensures that the largest value receives relatively greater weight; but as argued above, there are other reasons that Minkowski summation does not work well for the

streaks and bands application. The alternative local distortion-weighted pooling strategies investigated by Wang and Shang,⁸ are not suitable either, since in the limit of N identical defects the distortion-weighted pooling scales linearly with N . An alternative pooling method was proposed⁶ that attempts to more directly represent the observer's conscious pondering during stage 3 of the assessment:

$$M = \sum_{i=1}^N \frac{\tilde{\delta}_i}{p^{i-1}} \quad \text{with } \tilde{\delta}_i \geq \tilde{\delta}_{i+1}, \tag{7}$$

where M is the pooled value of the defect magnitudes, and is expected to correlate to the overall quality impairment. The parameter $p > 1$, controls the strength of the tent-pole effect, with larger values of p causing the larger defect magnitudes to dominate more. Note that as earlier, $\tilde{\Omega} = \{\tilde{\delta}_i\}_{i=1}^N$ is the set of visual defect magnitudes, but sorted in descending order of magnitude. Sorting the defects according to magnitude enables the tent-pole effect to be modeled more directly. When pooling according to Eq. (7) is performed on a series of defects with identical magnitude δ , the result is a geometric series that asymptotes to $\delta/(1 - 1/p)$. The cumulative tent-pole function for the pooling mechanism of Eq. (7) is shown in Fig. 4 for $p = 1.55$, which yields the best fit to the experimental data [if constrained to the form of Eq. (7)].

This discussion of the tent-pole function is based solely on considerations of images with multiple defects of identical visual magnitude. Without any further experimental evidence, we propose that the application of the tent-pole function can be generalized to also represent the diminishing significance assigned to defects in the presence of other defects with greater magnitude; that is, that the overall impairment is

$$\Delta Q(\tilde{\Omega}) = \sum_{i=1}^N T(i) \tilde{\delta}_i \quad \text{with } \tilde{\delta}_i \geq \tilde{\delta}_{i+1}. \tag{8}$$

We denote Eq. (8) as tent-pole pooling with tent-pole function T . The experimental data for the streaks and bands cumulative tent-pole function can be fitted well using the functional form

$$C_T(N) = 1 + a \tan^{-1}[(N-1)/b], \tag{9}$$

with $a = 1.33$ and $b = 3.0$, as shown in Fig. 4(b), from which $T(N)$ can be calculated. Note that the experimental data presented in Fig. 4(b) were based on a relatively small number of observers, and that the error bars could be significant, so that the form of Eq. (9) and its parameters should be taken to indicate only the general trend of the tent-pole function.

4 VBS Measurement Procedure

This section explains the VBS algorithm, in particular how the QIF and tent-pole pooling is applied for the measurement. In the Sec. 5, measurements using the VBS algorithm are compared with subjective assessments, using data obtained from the Japanese WG4 committee. The VBS algorithm development and implementation were completed

well before the WG4 activity started, so that this presents an opportunity to test the VBS algorithm on a set of diverse (class IV) images that did not influence the design and parameters of the algorithm. For that reason, this section describes exactly the implementation of the general concepts as was used for the WG4 data to be presented in Sec. 5, and it will be noted, that this implementation in some areas deviate slightly from the (perhaps more optimal) parameter choices indicated earlier in this paper. In most cases, these deviations are insignificant for the results. Note also that the method of separation into frequency bands, described shortly in step 6, leaves much to be desired, but does accurately describe the algorithm used for the WG4 data.

Although some of the steps have been discussed in the prior text, the entire procedure is summarized here, including some details of the procedure that was used for the measurements that are discussed in Sec. 5. The measurement requires a printed page that contains a region of at least 170×170 mm that is nominally uniform. In this section, a terminology is used where the 1-D defects (streaks or bands) run “vertically” such that a “horizontal” trace from left to right encounters the luminance variation caused by the defects. Although the algorithm could be applied to image regions that are much smaller than 170 mm in the vertical direction, that is, in practice, not recommended since 2-D defects, such as mottle, will more easily contaminate the measurement.

1. The print was scanned with an Epson Expression 10000XL flatbed scanner set to 600 dpi sampling frequency. *RGB* data were captured at 8 bits per channel.
2. Most of the print samples were printed in monochrome black only, but on some prints the gray color had been rendered with a mixture of *C*, *M*, *Y*, and *K*. A generic model to convert from *RGB* to CIE L^* had previously been derived for this scanner based on a large set of print samples with diverse materials and colorants. This generic model was used for all print samples to convert the scanned image values to L^* . Under more ideal circumstances, this step would use a conversion model that was optimized for the materials of the print.
3. The 2-D image was collapsed to a 1-D L^* profile, by averaging in the vertical direction. In practice, this used fiducial marks to compensate for skew in the placement of the print on the scanner.
4. The L^* profile was filtered using the QIF as given by Eq. (2) with parameters given in Table 2 for QIF 2, with the modification that frequencies above 0.5 c/mm are cut off to eliminate the influence of microuniformity defects on the measurement. The frequency cut-off was a requirement driven by the original purpose of the algorithm to replace a specific visual evaluation process and may not be desirable in general.^{6,13}
5. The profile $D(x)$ of the deviation from the mean was calculated, as in Eq. (3).
6. *Separation of profile into three frequency bands.* This step was not discussed previously in this paper, and deserves a few comments of justification. Real print samples often contain both wide and narrow streak

defects, which may well overlap each other. For example, a narrow light streak may be imposed in the middle of a wide dark streak. The human eye is very good at interpreting such a situation correctly, as two defects. But a straightforward analysis of the $D(x)$ profile could interpret this as two dark streaks (of half the width) and, if the amplitude of the light streak is large enough, as a third light streak. If the image was pooled with an L_p norm, as in Eq. (5), the distinction may not be severe, but when using tent-pole pooling it is important to account correctly for the individual defects. The profile $D(x)$ is separated into three profiles, corresponding to high, medium, and low spatial frequencies, using three normalized Gaussian convolution kernels as follows.

$$G_i(x) = \frac{1}{(\pi w_i^2)^{1/2}} \exp[-(x/w_i)^2],$$

with $w_1 = 50$ mm, $w_2 = 5$ mm,
 $w_3 = 0.5$ mm, (10)

$$\begin{cases} D_1(x) = G_1 * D \\ D_2(x) = G_2 * D - G_1 * D \\ D_3(x) = G_3 * D - G_2 * D, \end{cases} \quad (11)$$

where $*$ denotes the convolution operator. As already mentioned above, the separation process is problematic, for example, it may generate spurious peaks in the separated signals. A method of cleanly separating overlapping defects would be desirable.

7. From each of the three profiles the extrema were identified and their absolute values were interpreted as defect magnitudes $\{\delta_{ki}\}_{i=1}^{N_k}$, where the index $k=1,2,3$ identifies the profile. The defect magnitudes are sorted into $\{\tilde{\delta}_i\}_{i=1}^N$, $\tilde{\delta}_i > \tilde{\delta}_{i+1}$, where $N=N_1+N_2+N_3$ is the total number of defects from all three profiles. A threshold value of 0.05 was subtracted[†] from the defect magnitudes (and if this led to a negative value, the magnitude was set to 0).
8. The defect magnitudes $\{\tilde{\delta}_i\}_{i=1}^N$ were pooled using Eq. (7) with parameter $p=2$ into the overall impairment[‡] M .
9. A nonlinear scaling is performed: $VBS = 3.66\sqrt{M}$ (the factor 3.66 is arbitrary and for historical reasons only).

5 Correlation to Subjective Assessments

This section presents a comparison of predictions by the VBS algorithm with subjective assessments of print samples obtained from actual printers, including a mix of electrophotographic (EP) and ink jet (IJ) technologies. The print samples and results of the subjective analysis were kindly provided by the SC28/WG4 committee of Japan. The Japanese WG4 committee took charge of generation of

[†]This is a slight deviation from the general concept, with minor impact. The value 0.05 was chosen based on experts' examination of print samples that appeared perfect.

[‡]This is also a slight deviation from the more optimal use of Eq. (9) with $p=1.55$.

Table 3 Overview of the characteristics of the sample.

Group	H	V	Invisible	Mixed Defects		Total Sample Count
				Both	2-D	
Vertical	—	20	6	—	—	26
Horizontal	2	—	6	—	—	8
IJ	—	15	5	—	—	20
EP	2	5	1	6	1	15
Mixed	—	—	—	6	1	7

The first column shows the grouping used for the analysis in Sec. 5.2. Columns 2–6 correspond to the tags assigned by the WG4 committee according to the visibly dominating defect. The last column shows total number of samples in the group. For example, the samples in the 2nd row (Vertical) are those that can be expected to correlate well with vertical VBS measurements, which includes both those samples with tagged “V” and those with nearly invisible defects.

the print samples for an experiment that verifies image quality attributes for ISO/IEC 24790 development. The SC28/WG4 committee of Japan, the United States, the Netherlands, South Korea, and China participated in the subjective evaluations. The observers were predominantly image quality experts and trained observers.

The sample set consisted of 35 prints (originally 38, but 3 prints were omitted by the WG4 committee due to irregularities with the prints and/or their assessment). The sample set was diverse, including samples dominated by streaks, as well as samples dominated by periodic banding. The samples had been tagged by the Japanese WG4 committee according to marking technology and the nature of the visibly dominating defects on the prints, as follows: “V” = 1-D defects (streaks or bands) in the vertical direction, “H” = 1-D defects (streaks or bands) in the horizontal direction, “Both” = 1-D defects in both vertical and horizontal directions, “2D” = random 2-D noise such as mottle, and “Invisible” = no significant visible defects. These tags were unknown to the observers, and not used until the data analysis. Table 3 gives an overview of the sample characteristics, and explains how the samples were grouped for the analyses described in Sec. 5.2.

5.1 Subjective Methods

Two subjective assessment methods were used: categorical scaling and anchored scaling. The categorical scaling used five categories: (1) “Banding is imperceptible,” (2) “Banding is slightly perceptible,” (3) “Banding is perceptible but not annoying,” (4) “Banding is annoying,” (5) “Banding is very annoying.” The anchored scaling used a print with low levels of defects as anchor with scale value 25, and a print with high levels of defects as anchor with scale value 80. In both cases, observers were asked to assess the overall quality with respect to streaks and bands. The correlation between the results of the two methods was very high, and the small differences would not affect the conclusions of this paper; therefore, this section presents only the data from one of the methods—the anchored scaling method. However, the categories just listed give the reader a good sense of the range of quality of the print samples. Note that the

anchored scale value of 25 corresponds to category 1, while anchored scale value of 80 corresponds to category 5.

5.2 Comparison of VBS to Subjective Results

Real print samples will typically contain streaks and bands defects in both directions, as well as 2-D noise such as mottle. This complicates the subjective assessment in the case where the observer is asked to assess only streaks and bands. First, if mottle levels are sufficiently high relative to the streaks and bands, then mottle may mask the perception of streaks and bands. Second, the interaction of vertical and horizontal 1-D defects may give an appearance of two-dimensionality, which could lead the observer unintentionally to disregard some of the streaks and bands.

From a measurement perspective the presence of these multiple defects is less of an issue because, as mentioned earlier, when large image regions are used for the analysis, typical mottle defects will not significantly impact the profile average. However, for prints that contain both horizontal and vertical 1-D defects, the question of how the measures for each direction are combined into a single assessment, is outside the scope of the VBS algorithm.

To analyze the data, the samples were grouped as shown in Table 3. The “Vertical” group contains those samples that do not have visibly significant defects other than in the vertical direction, including those samples that do not have visibly significant defects at all. The samples in this group can meaningfully be analyzed by application of the VBS algorithm in one direction. Similarly for the “Horizontal” group, but in this case the VBS algorithm is applied in the other direction. Figure 5(a) shows the performance of the VBS algorithm on these two groups. Note that the samples tagged “Invisible” are included twice in this graph, once for each VBS direction. The summary statistics for a linear regression between VBS and the subjective ratings are given in Table 4. Spearman’s rank order correlation is 0.93, and the relationship is nearly linear with an adjusted R^2 of 0.87. Given that this sample set has defect diversity of class IV, the performance is quite satisfactory. Note also that some noise is introduced, since the VBS data measures

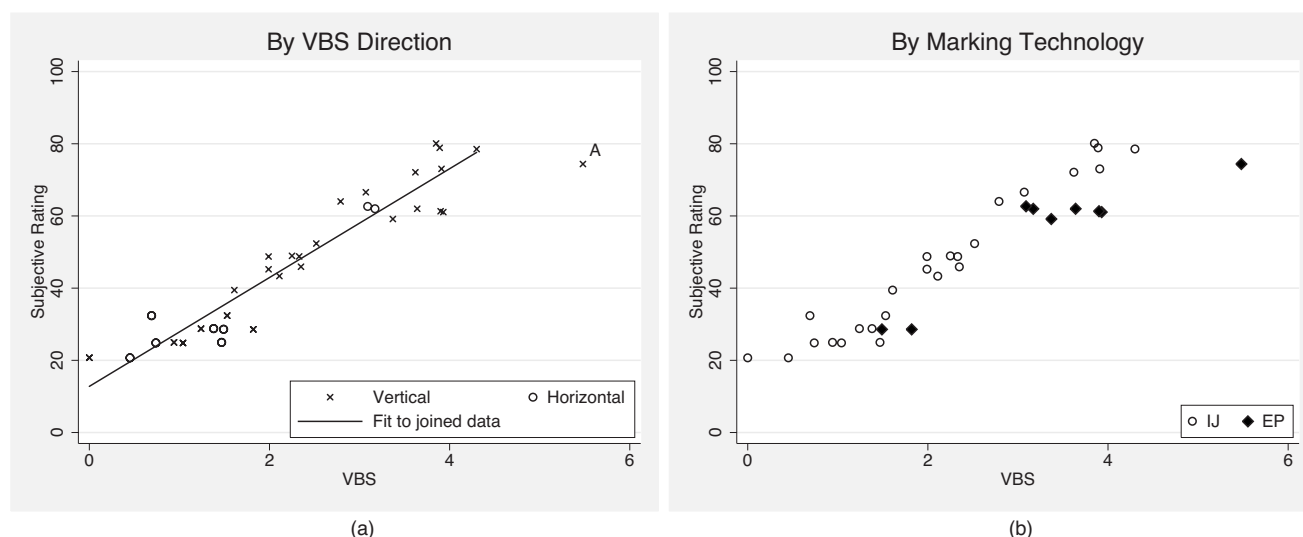


Fig. 5 Subjective ratings versus VBS for samples dominated by 1-D defects in a single direction, including samples with no significant defects at all: (a) the line is fitted with the outlier “A” removed and (b) the same as (a), but stratified by marking technology. The subjective data are not available in a form that enables rigorous calculation of error bars; however, comparison of results obtained from five different observer groups indicate an agreement between group averages within ± 5 to 7 units.

only one direction (the one presumed to dominate), while the subjective assessment takes both directions into account. There is perhaps one data point that could be considered an outlier: The point labeled “A” (with the largest VBS value). This point received a subjective rating close to the maximum anchor value, and even though observers were allowed to give ratings beyond the anchor values, it can be expected that there is some resistance for the observers to do so. Omitting this point from the regression, changes the adjusted R^2 and Spearman’s rank correlation only slightly, but reduces the root mean squared error (MSE) from 6.9 to 6.1. Figure 5(b) shows the same data stratified by the marking technology, and regression results for each technology are given in Table 4. Note that the correlation between VBS and subjective ratings is particularly good for the ink jet technology samples.

As seen from Table 3 there are seven samples in the “mixed” group. One of these has 2-D defects, such as mottle, while the other six have streaks and/or bands in both the vertical and horizontal direction. Figure 6 shows the subjective ratings versus VBS for these samples. For each sample, VBS is evaluated twice, once in each direction. Compared to the regression from the pure 1-D defects of Fig. 5, these data show a clear tendency for the VBS ratings to be higher (worse) than the subjective ratings. This tendency is the opposite of what could be expected, given that VBS is supposed to be rather insensitive to 2-D defects, and given that for each data point VBS is evaluated only in one of the two directions, while the subjective ratings take both directions into account simultaneously. The explanation for this behavior is not known at this point, and it may require a larger sample set to reach an understanding

Table 4 Summary statistics for the prediction of subjective ratings by VBS, as shown in Fig. 5, with the 95% confidence intervals given for the intercept and slope.

	<i>N</i>	Intercept	Slope	Root MSE	Adjusted R^2	Spearman’s rank correlation
Streaks/bands in single direction (H or V)	34	15.0 [9.9, 20.1]	13.9 (12.0, 15.8)	6.9	0.87	0.93
Streaks/bands in single direction (H or V), one outlier removed	33	12.8 [8.0, 17.6]	15.1 [13.2, 16.9]	6.1	0.90	0.92
By marking (IJ)	25	12.0 [7.9, 16.2]	16.3 [14.6, 18.0]	4.9	0.94	0.96
By marking (EP)	9	14.8 [-2.5, 32.1]	12.3 [7.3, 17.2]	7.0	0.81	0.54

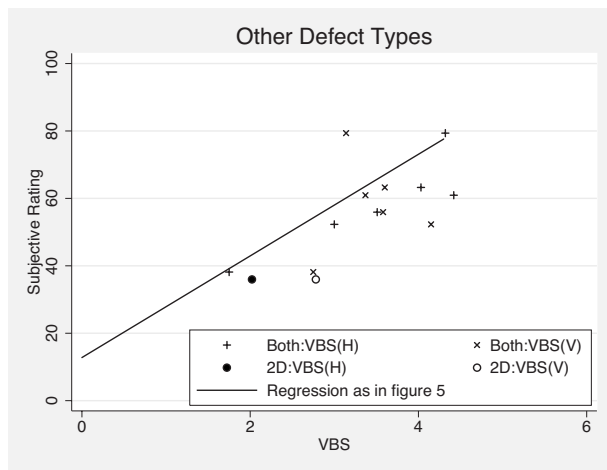


Fig. 6 Subjective ratings versus VBS for samples with defects of a mixed type, not purely one dimensional, whose ratings are not expected to be predicted well by VBS.

of this issue. Potential explanations may involve masking, as described at the beginning of this section.

6 Discussion

The philosophy behind the algorithm discussed in this paper is to directly address each of the three important stages of human visual assessment of streaks and bands: (1) construction of the perceived image, (2) identification and quantification of individual defects in the perceived image, and (3) the conscious integration of the set of defects into a single numerical value that represents the overall quality of the print. In reality, the VBS algorithm has several obvious limitations and to some extent violates the distinction between the three stages. To construct the perceived image the VBS algorithm uses a simple CSF-like filter, which ignores more complex aspects of the human visual system; for example, it ignores masking caused by 2-D noise in the print. Furthermore, to achieve a clear-cut way of assigning a visual magnitude to each defect, VBS uses a QIF rather than a CSF, which means that the filtered image cannot necessarily be interpreted as the “perceived image.” In practice, however, there is not any one CSF that suits all viewing conditions, and the QIF is nothing other than a CSF tuned for this particular purpose. When prints are viewed at normal viewing distance of about 30 to 40 cm, an edge-to-edge gradient of 1 L^* difference is barely perceptible, while 2-mm periodic banding is visible at much lower amplitudes. Therefore, for streaks and bands assessment it is critical that the QIF maintains the bandpass nature of the CSF, where low spatial frequencies are strongly suppressed. Other than that, the QIF will depend on numerous details, for example on the size of the inspected image region, as illustrated by Fig. 3.

As for stage 2, the merit of the algorithm is that it uses stage 1 with the QIF to provide a logical foundation for how to assign “visual magnitudes” to each individual defect. One obvious limitation of the algorithm during stage 2, is that the separation of the L^* profile into several frequency bands will not be able to accurately capture all defect configurations and, as mentioned, in some cases can lead to spurious defect identifications.

Finally for stage 3, it is noteworthy that the relationship between VBS and the subjective rating is nearly independent of whether the print is generated by IJ technology (prone to streaks) or EP (prone to bands), which may be a sign that the defect pooling is capturing some of the right effects of the human quality judgment. It is indeed a fundamental assumption of the VBS approach that the impact of periodic defects such as bands can be modeled as unrelated streaklike defects, and it would be desirable in future work to test this assumption directly. It is important to understand that the tent-pole defect pooling does not represent the (subconscious) visual masking of defects by each other. Rather, it aims to represent a conscious “mental masking” during the judgment process. In reality, there may be a gray area between visual masking and mental masking, but there are cases where the masking is clearly mental. The experiment to determine the tent-pole function was quite limited in scope. In particular, it addressed directly only the case where multiple defects of identical magnitude were being pooled; furthermore, the number of observers was so small that meaningful confidence intervals were not determined. It would be worthwhile to conduct experiments to probe pooling of streaks and bands defects in greater detail, including the more realistic situation where defect amplitudes are not identical.

The VBS algorithm is sensitive only to L^* variations, but based on the principles already described, it would be straightforward to extend the algorithm to be able to analyze full-color variations. For many printing systems, streaks and bands are usually associated with L^* variations that tend to dominate the perceived defects, in which case VBS is sufficient, but situations do arise where sensitivity to a^* and b^* is necessary, for example, in cases of low-frequency banding with slight hue changes.

The use of the human CSF in algorithms for print quality is by no means new. The slight variation introduced by the QIF, as opposed to the CSF, may also be relatively insignificant, compared to the variations seen in the CSF as a function of viewing conditions. Thus, the question is which part of the VBS algorithm is most significant for obtaining reasonable agreement with subjective assessments. It would be desirable for future work to explore the relative significance of, for example, the QIF and the tent-pole pooling method. To do so will require significantly larger image sets and observer pools, and therefore the experiments would be more suitable for softcopy than hard-copy assessments.

As a final critique of this approach to assessment of printer image quality, it should be considered to what extent the measurement is relevant, even if it did correlate perfectly with the subjective assessments of the test charts. This question leads back to stage 3 of the observer’s judgment of overall quality. How does an observer judge the overall quality of a nominally uniform page, containing streaks and bands defects—what are the value perceptions that form the basis for the assessment? Without a context, such as a real customer document in which to evaluate the quality, it is difficult to perform, or at least to understand, the evaluation. For this reason, some of the subjective experiments described in this paper attempted to give a context by directing the observer to imagine that the image would be used for a document cover page. Nevertheless,

printer manufacturers are in strong agreement that it is important to establish methods and standards for this kind of attribute assessment, and it is clear from numerous subjective experiments, including those by the SC28/WG4 committee, that such assessments are fairly precise and reproducible.

Acknowledgments

The author would like to acknowledge the contributions from the ISO/IEC SC28/WG4 committee of Japan for generating and providing print samples and subjective analysis results, and would like to acknowledge the contributions from the SC28/WG4 committee of Japan, the United States, the Netherlands, South Korea, and China for contributing to the subjective evaluations.

References

1. H. Mizes, N. Goodman, and P. Butterfield, "The perceptibility of random streaking," in *Proc. IS&T PICS*, pp. 89–93, Portland, OR (2000).
2. C. Cui, D. Cao, and S. Love, "Measuring visual threshold of inkjet banding," in *Proc. IS&T PICS*, pp. 84–89, Montreal, Canada (2001).
3. O. Arslan, Z. Pizlo, and J. P. Allebach, "Softcopy banding visibility assessment," in *Electronic Imaging, Image Quality and System Performance, Proc. SPIE 5294*, 38–50 (2004).
4. D. R. Rasmussen, K. D. Donohue, Y. S. Ng, W. C. Kress, F. Gaykema, and S. Zoltner, "ISO 19751 macro-uniformity," in *Electronic Imaging, Image Quality and System Performance, Proc. SPIE 6059*, 60590K (2006).
5. D. R. Rasmussen, P. A. Crean, F. Nakaya, M. Sato, and E. N. Dalal, "Image quality metrics: applications and requirements," in *Proc. IS&T PICS*, pp. 174–178, Portland, OR (1998).
6. D. R. Rasmussen, E. N. Dalal, and K. M. Hoffman, "Measurement of macro-uniformity: streaks, bands, mottle, and chromatic variations," in *Proc. IS&T PICS*, pp. 90–95, Montréal, Canada (2001).
7. Z. Wang and X. Shang, "Spatial pooling strategies for perceptual image quality assessment," in *Proc. IEEE Inter. Conf. Image Proceeding*, pp. 2945–2948, Atlanta, GA (2006).
8. J. C. Briggs, M. Murphy, and Y. Pan, "Banding characterization for inkjet printing," in *Proc. IS&T PICS*, pp. 84–88, Portland, OR (2000).
9. A. Eid, B. Cooper, and E. Rippetoe, "A unified framework for physical print quality," in *Electronic Imaging, Image Quality and System Performance, Proc. SPIE 6494*, 64940C (2007).
10. K. D. Donohue, C. Cui, and M. V. Venkatesh, "Wavelet analysis of print defects," in *Proc. IS&T PICS*, pp. 42–47, Portland, OR (2002).
11. P. J. Kane, T. F. Bouk, P. D. Burns, and A. D. Thompson, "Quantification of banding, streaking, and grain in flat field images," in *Proc. IS&T PICS*, pp. 79–83, Portland, OR (2000).
12. Y. Bang, Z. Pizlo, J. P. Allebach, and N. Burningham, "Discrimination based banding assessment," in *Proc. IS&T NIP19*, pp. 745–750 (2003).
13. E. N. Dalal, D. R. Rasmussen, F. Nakaya, P. A. Crean, and M. Sato, "Evaluating the overall image quality of hardcopy output," in *Proc. IS&T PICS*, pp. 169–173, Portland, OR (1998).
14. D. Marr, *Vision*, W. H. Freeman and Company, New York (1982).
15. S. E. Palmer, *Vision Science*, The MIT Press, Cambridge, MA (1999).
16. P. G. J. Barten, *Contrast Sensitivity of the Human Eye and Its Effects on Image Quality*, SPIE Publications, Bellingham, WA (1999).
17. J. C. Dainty and R. Shaw, *Image Science*, Academic Press, San Diego, CA (1974).
18. R. P. Dooley and R. Shaw, "Noise perception in electrophotography," *J. Appl. Photogr. Eng.* **5**(4), 190–196 (1979).
19. K. Kagitani, M. Hino, and S. Imakawa, "Image noise evaluation method for color hardcopy," in *Proc. IS&T NIP12*, pp. 173–176 (1996).
20. J. Lubin, "The use of psychophysical data and models in the analysis of display system performance," Chap. 13 in *Digital Images and Human Vision*, A. B. Watson, Ed., pp. 163–178, The MIT Press, Cambridge, MA (1993).
21. S. Daly, "The visible differences predictor: an algorithm for the assessment of image fidelity," Chap. 14 in *Digital Images and Human Vision*, A. B. Watson, Ed., pp. 179–206, The MIT Press, Cambridge, MA (1993).
22. C. Taylor, Z. Pizlo, J. P. Allebach, and C. A. Bouman, "Image quality assessment with a Gabor pyramidal model of the human visual system," in *Electronic Imaging, Human Vision and Electronic Imaging, Proc. SPIE 3016*, 58–69 (1997).
23. J. G. Robson and N. V. Graham, "Probability summation and regional variation in contrast sensitivity across the visual field," *Vision Res.* **21**, 409–418 (1981).
24. M. To, P. G. Lovell, T. Troscianko, and D. Tolhurst, "Summation of perceptual cues in natural visual scenes," *Proc. Biol. Sci.* **275**(1649), 2299–2308 (2008).



D. René Rasmussen received his Lic Scient/PhD degree in physics from the Niels Bohr Institute, University of Copenhagen, in 1990, spending a couple of years as a visiting scientist at Cornell University. In 1992 he joined Xerox Corporation, where he is now principal scientist, technical manager of the Device Imaging and System Quality Optimization area of Xerox Innovation Group, and program manager for research on image quality and system color controls. His major technical interests are in the field of image quality analysis. He leads the International Committee for Information Technology Standards (INCITS) W1.1 macrouniformity team, and is member of the IS&T and SPIE.