Print Quality Metrics for Grayscale Text

Joyce E. Farrell
Hewlett-Packard Laboratories
Palo Alto, California

Abstract

The effect of printer resolution and grayscale upon text print quality was evaluated both empirically and analytically. Empirical data on subjective print quality was obtained in two ways. First, people were asked to rate print quality on a scale of 1 to 10 where 10 was considered to be optimal. Second, people were presented with pairwise presentations of different print samples and asked to indicate which sample they preferred. Both methods indicate that text print quality increases with printer resolution and number of gray levels.

The relationship between subjective ratings of text print quality and several attribute-based and image-based machine vision metrics was evaluated. Subjective print quality ratings are monotonically related to measurements of character jaggedness and unrelated to measurements of edge sharpness. Subjective ratings are also monotonically related to an image-based metric that attempts to quantify the perceived difference between any particular print sample and an ideal high-resolution print sample (approximated, in this case, by an offset print sample). The metric is computed by filtering the difference between digitized images of a character printed with offset technology and the same character printed with inkjet technology using published contrast sensitivity functions. Empirical print quality ratings are monotonically related to the visually-weighted difference between an offset print sample and the lower-resolution inkjet print samples.

Introduction

Machine vision metrics that reliably predict how a diversified population will judge print quality is a challenging, if not illusive, goal. In this paper we evaluate the relationship between subjective ratings of print quality and two types of machine vision metrics: attribute and image-based metrics. Attribute metrics are measurements of identifiable features in print samples that are known to influence subjective print quality. For example, optical ink density, character jaggedness and character edge sharpness are three very important attributes that have been shown to affect text print quality [1]. Image-based metrics are based on measurements of the the entire image (i.e. gestalt) of a character. The image-based metric we evaluate in this paper is designed to capture the perceptual similarity between a test print sample and a standard offset print sample by quantifying how 'close' a particular printed character is to the same character printed with offset lithography.

The print samples evaluated in this study were generated by filtering and sampling high-resolution binary (black and white) master characters. The sampling resolutions investigated were 150, 180 and 300 dpi. At all three resolutions, binary and grayscale versions of text characters were created by quantizing the filtered and sampled characters. Filtering the character with appropriate convolution kernels, effectively blurs the edges of characters, replacing aliasing errors (perceived as 'jaggies') with less-objectionable errors – namely, edge blurring.

Empirical Evaluation

Stimuli

Outline descriptions of each alphanumeric character in a Times Roman Font were used to generate high resolution (256 x 256) bi-level master characters. Different filtered and sampled versions of each alphanumeric character were created by convolving the bi-level master character with a box filter, sampling the filtered image with a lower-resolution sampling grid and then varying the intensity quantization levels of character pixel intensities.

Figure 1 shows a sampling grid superimposed upon a high-resolution master character. Each box region in Figure 1 represents a single sample point whose intensity is determined by the average of all the points in the the high-resolution image that fall within the box area. The intensity of each low-resolution sample or pixel is thus the unweighted average of image points within a specifed area. We refer to the box averaging as a box filter convolution. The filter support is the area over which the box filter convolution kernel is defined, as illustrated in Figure 1. We created grayscale characters using a filter support such that the image points that were averaged to generate the intensity of each pixel did not overlap, but rather abutted, as in Figure 1.

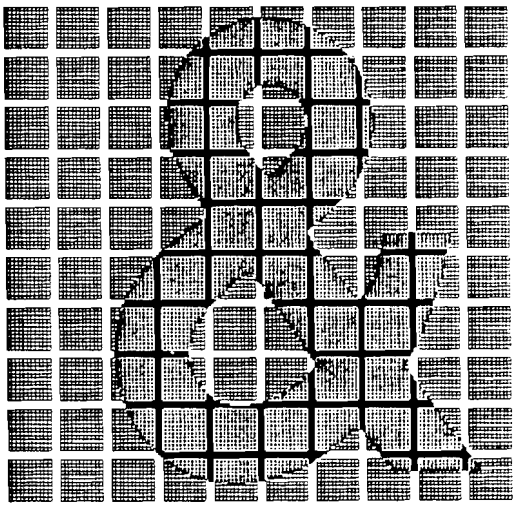

Figure 1. Illustration of sampling grid superimposed upon a high-resolution master character. The intensity of each sample point is computed by averaging all image points that fall within a square area.

Different sampled and filtered characters were generated for printer resolutions of 150, 180, 300 and 1280 dpi. To create binary print samples for each printer resolution, the sampled and filtered characters were quantized to two levels. The 1280 dpi binary sample was printed on a linotron printer and reproduced on coated stock paper with offset lithography. To create the grayscale characters, intensities were quantized to 12 levels for 150 dpi, 7 levels for 180 dpi and 4 levels for 300 dpi. These levels were then assigned to different dot sizes at each printer resolution. The assignment of gray values to dot sizes was achieved by trial and error and represents our best attempt to optimize the grayscale sampling and at the same time match apparent contrast across samples (see discussion of apparent contrast below).

The stimuli consisted of a text passage from *Huckleberry Finn* by Mark Twain and several lines of characters illustrating the entire upper and lower-case alphabet. Print samples were designed to differ only in grayscale and resolution. Text characters were matched in size, font and apparent contrast. The text passage was printed with 6 pt. Times Roman font characters. The nominal height of capital alphanumeric characters (estimated by the height of the character 'W') was 2.169 mm and the nominal width of capital characters was 1.595 mm. Special care was taken to match the horizontal and vertical spacing of characters in the text passages. Across all print samples, character leading (vertical distance between lines) was approximately 2.26 mm and character spacing (horizontal line width/number of character spaces) was approximately 0.8484 mm.

The text was printed by an experimental multi-drop inkjet printhead onto coated stock paper with paper reflectance of approximately 87%. All print samples were printed using a single inkjet head that constrained the minimum dot size one could print at 300 dpi to be 0.05 mm. Due to this constraint, we were limited to 3 different distinquishable dot sizes at 300 dpi, 7 different dot sizes at 180 dpi and 12 different dot sizes at 150 dpi. Although the number of achievable graylevels varies with printer resolution, the dynamic range of the ink densities was matched by selecting inkjet dot sizes such that the maximum ink density was approximately 1.6 across all print samples

Procedure

Subjects were presented with a total of 7 print samples: 150 dpi with 2 and 12 levels of gray, 180 dpi with 2 and 7 levels of gray, 300 dpi with 2 and 4 levels of gray and the offset print sample. Print samples were illuminated with a 32 watt, 11 inch diameter Phillips cool-white fluorescent ring lamp. This lamp produced an illumination of approximately 2387.61 lux (221.92 foot-candles) which is comparable to daylight out-of-doors or good interior lighting (Boynton, 1966). Under this lighting, the paper luminance was $640 cd/m^2$. Subjects viewed the print samples from a viewing distance of 10 inches.

Subjective print quality was assessed using two methods: a rating method of magnitude estimation and a preference judgement task. The viewing conditions for the two tasks were identical. In the rating task, subjects were shown an offset print sample of the text passage printed in the 6 pt Times Roman font. Subjects then viewed each inkjet print sample and rated the print sample on a scale of 1 to 10 where 10 represented optimal print quality defined by the offset print sample. Subjects viewed 10 presentations of each of the 6 inkjet print samples, presented in a random order of presentation. Eight subjects participated in the rating task. Four of the subjects (referred to as 'experts') were members of the HP Labs Printing Development Laboratory where evaluating print quality was an integral part of their work. The remaining four subjects ('novices') had no previous experience in print quality evaluation. Two subjects participated in a preference judgement task. They viewed all possible pairwise presentations of the seven print samples, including the offset print sample. For each pairwise comparison, subjects indicated which print sample had the better image quality. Over the course of several days, the two subjects viewed a total of 420 trials arranged in a random order of presentation.

Results

We assume that subjective print quality can be represented by a single number that is unique up to a monotonic scale transformation. In other words, we assume subject's ratings of print quality can be used to order print samples from low to high print quality or from 'least preferred' to 'most preferred'. We tested this assumption by looking for violations of transitivity in subjects' preference judgements: For example, if subjects preferred sample a over sample b and preferred sample b over sample c, then subjects should have also preferred sample a over sample c. Violations of transitivity in subjects' preference judgements would indicate that an ordinal scale for print quality did not exist. There were no violations of transitivity for both subjects who participated in the preference judgement task.

A preference weight is estimated for each print sample by the number of times that each print sample was preferred over all other print samples in the preference judgement task. The rank-order correlation between preference weights and print quality ratings is 1.0. These results support the assumption that subjects' ratings can be interpreted with respect to an ordinal scale for print quality.

The averaged ratings for novices and experts in the rating task did not differ significantly and were, therefore, averaged. The print quality ratings shown in Figure 2 represent the ratings averaged across the eight subjects who participated in the rating task. Figure 2 shows that the mean print quality ratings increase with printer resolution and with the introduction of grayscale filtering. Figure 2 also illustrates the trade-off between grayscale and printer resolution. For example, 150 dpi with 12 levels of gray is rated as having comparable print quality as the 180 dpi binary print sample: The print quality ratings for these two print samples do not differ significantly by a Wilcoxen sign test with $p \leq .05$. Clearly at all printer resolutions (150, 180 and 300 dpi), the introduction of grayscale improved the subjective image quality of the printed text. Note, however, that because we did not vary the number of graylevels for each printer resolution, we cannot assume that print quality will continue to increase as the number of graylevels increases.

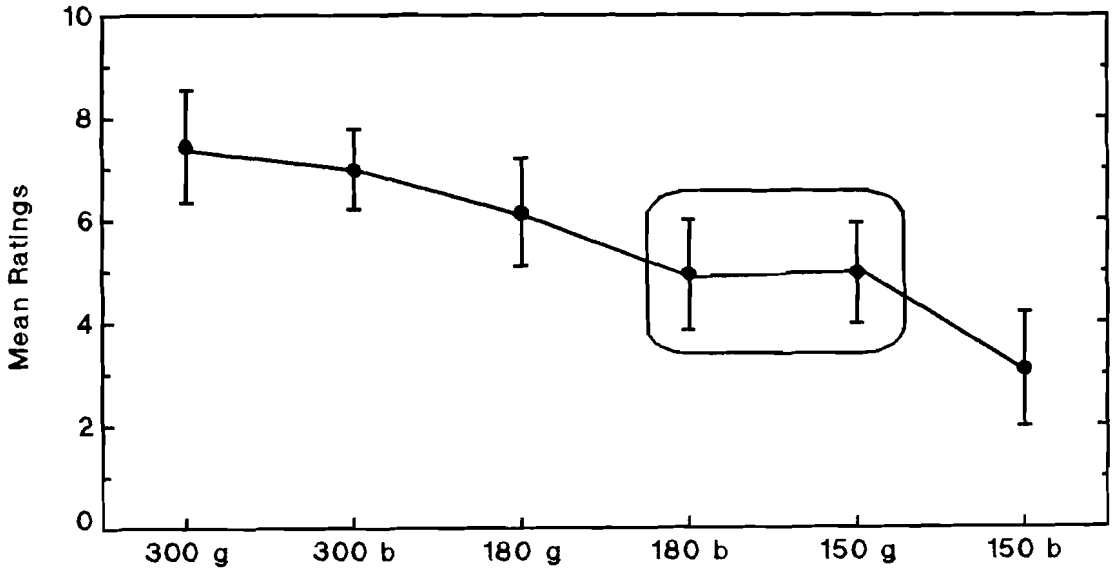

Figure 2. Mean print quality ratings for each inkjet print sample plotted for binary and grayscale characters at each printer resolution.

Analytical Evaluation

Text characters were digitized using a General Electric CID camera equiped with 33 mm macroscopic lens and attached to a Poynting frame grabber. Digitized images of the test character had a sampling resolution of 512 x 512 and 8 bits of intensity. Each sampled point in the digitized image was separated by 10 microns. The illumination of the print samples was comparable to the illumination of the print samples in the print viewing station used for the subjective assessment of print quality. Print samples were illuminated with approximately 2356.19 lux or 219 foot-candles. The test character selected for our initial analysis was the capital letter 'W'. This character was clearly affected by the filtering technique used to generate the grayscale characters and was, consequently, diagnostic of the improvements in print quality due to the introduction of grayscale.

Attribute metrics

Grayscale characters pose a challenge for attribute metrics that predict that subjective print quality increases with edge sharpness and decreases with edge jaggedness. Grayscale characters reduce edge jaggedness by decreasing edge sharpness. Thus, we would expect that subjective text print quality would be inversely related to edge jaggedness, but not necessarily related to edge sharpness. By holding the maximum optical density constant across all the print samples, we were able to investigate the relationship between grayscale filtering, edge jaggedness and edge sharpness.

Since the size of characters was held constant across all print samples, character contour jaggedness can be estimated by the perimeter of a character. We measured inner and outer perimeters of the digitized image of the test character by following the contour of a designated grayscale value in the digitized image of the test character. The low and high grayscale values were selected on the basis of a histogram analysis of all digitized grayscale versions of the test character. The character contour jaggedness for each print sample was estimated by the average of the inner and outer perimeters measured from the digitized printed test character. One of the limitations of this metric is that it is dependent on the particular gray values selected for the inner and outer perimeters.

Edge sharpness was estimated by dividing the the area of the character's edge (estimated by the area between the inner and outer perimeters) by the average perimeter. The edge sharpness metric is, then, an estimate of the average width of the transition from high to low gray levels at the character's edge. If the width is small then the distance between the high and low gray values at the character's edge is small or, in other words, the slope of the contrast change at the character's edge is high. Conversely, if the average transition width is large, the slope of the contrast change is low.

Figure 3 shows that print quality rating and preference judgements are monotonically related to the character jaggedness measure. The 180 dpi binary print sample is the only print sample that violates this monotonic relationship. The rank-order correlation between print quality ratings and character jaggedness is -0.94 which is significant with $p \leq .05$. Moreover, the relationship between print quality ratings and character jaggedness is consistent with the effect we would expect character jaggedness to have on print quality: The smaller character jaggedness measure the better the print quality.

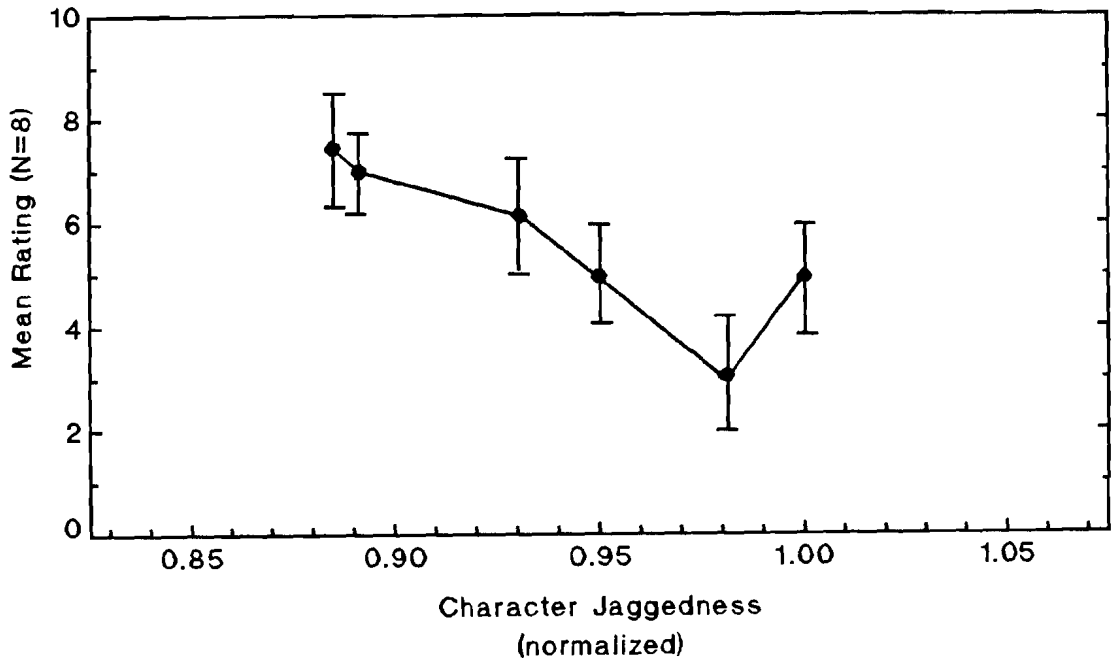

Figure 3. Mean print quality ratings for each inkjet print sample plotted a function of the normalized character jaggedness metric.

Figure 4 shows that print quality ratings and preference judgements are not monotonically related to the measurement of edge sharpness. Four print samples violate a monotonic relationship. The rank-order correlation between print quality ratings and character edge sharpness is -0.83 which is not significant with $p \leq .05$. This result leads us to conclude that there is no relationship between the observed improvement in print quality with printer resolution and grayscale and the measurement of edge sharpness.

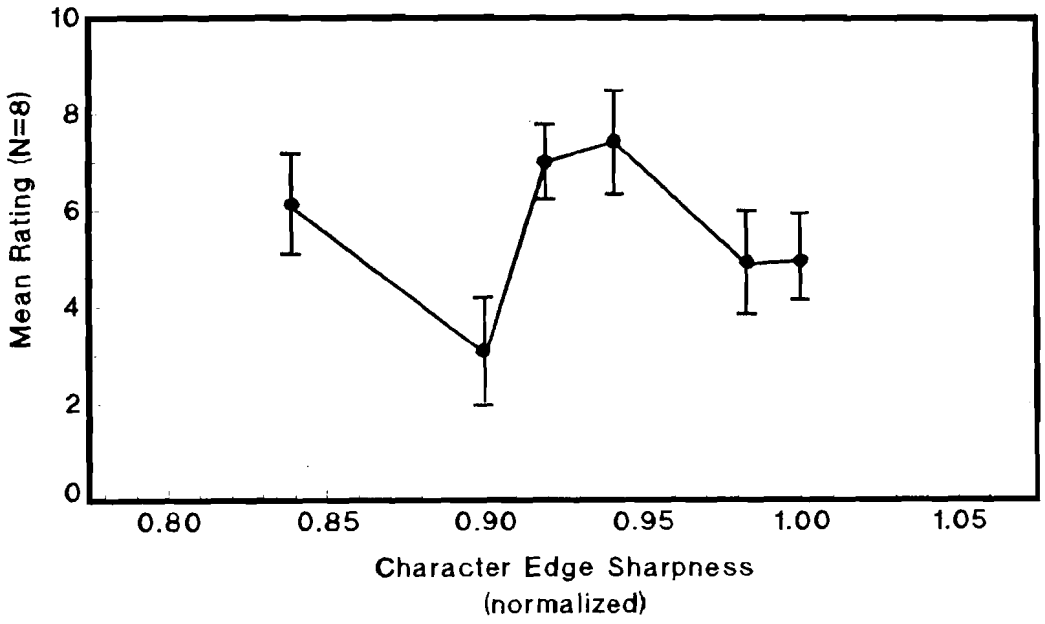

Figure 4. Mean print quality ratings for each inkjet print sample plotted a function of the normalized edge sharpness metric.

There are several reasons to expect edge sharpness to be a poor predictor of print quality. First, edge sharpness does not capture the improvement of print quality with resolution: With the exception of the offset print sample, the low-resolution characters do not have larger edge transition widths than the higher resolution characters. Second, grayscale characters violate the the expectation that characters with sharp edges have high print quality: The average transition width increases when grayscale is introduced (since the effect of grayscale filtering is to blur the contrast at a character's edge) and yet all subjects rated the grayscale characters with blurred edges to have higher print quality than the binary characters with sharp edges.

The fact that subjects' ratings of print quality are related to character jaggedness but unrelated to character edge sharpness supports Hamerly's conclusion [2] that jaggedness will be the limiting image-quality factor for relatively low-resolution imaging devices, such as inkjet and laser printers, whereas edge sharpness will be the limiting factor for high resolution imaging devices, such as continuous-tone silver halide photography.

The Visually-Weighted Difference Metric

The visually-weighted difference between any two print samples is calculated by taking the difference between the digitized images of the print samples, filtering the difference with an appropriate constrast sensitivity function, and summing the squared result [3]. The metric is based on several assumptions. First, we assume that ink/paper reflectances that differ in equal steps on the Munsell scale for lightness are perceived to be equally separated in perceived brightness [4]. (Graylevel intensities of the digitized images of test characters were converted to Munsell values using a gamma look-up table relating the camera's response to calibrated gray paper chips that sampled the entire range of the Munsell scale of lightness.) Second, we assume that human spatial contrast sensitivity can be described by a single contrast sensitivity function of the form

$$C(f) = K(bf)^a e^{-bf} \tag{1}$$

where a and b are parameters that depend on the ambient light, pupil size or any other visual factors that may change the state of visual adaptation [5,6].

To approximate human contrast sensitivity for equivalent viewing conditions, we derived parameter values by fitting Equation (1) (above) to Campbell's (1968) measurements of the contrast sensitivity function for a 3.8 mm pupil [7]. Of the many contrast sensitivity functions published in the literature, we considered this function to be appropriate on the grounds that the pupil size, estimated from the the DC stimulus luminance in our experiments, is approximately 4 mm. [8].

To quantify the difference between two visually filtered or weighted images, we compute the vector length of the filtered difference between the two images. The vector length is based on the Euclidean distance between two points in an n-dimensional space,

$$\sqrt{\sum_{i=1}^{n}(A_i - B_i)^2} \tag{2}$$

where A_i and B_i are corresponding points in images A and B, respectively. For computational reasons, we use the square of the Euclidean distance or, the squared vector length of the difference.

The visually-weighted difference metric was originally designed to predict when two images will appear to be different [3]. To extend the metric to image quality, we assume that subjective print quality is monotonically related to the perceived similarity between the print sample and an idealized standard. In other words, the image quality of a particular print sample should be inversely related to the visually-weighted difference between the sample and an 'ideal' high-resolution print sample: The smaller the visually-weighted difference, the greater the perceived image quality. We test this assumption by comparing the visually-weighted difference (between text printed on inkjet printers varying in resolution and grayscale capability and the same text printed on the same paper with offset lithography) to subjective print quality ratings.

Figure 5 shows the print quality ratings for each inkjet print sample plotted as a function of the visually-weighted difference between each inkjet print sample and the offset print sample. The visually-weighted differences plotted in Figure 5 is based on an analysis of the 'W' character. Figure 6 shows the results of an analysis of the '/' character. Both figures show a similar trend: With one exception (see Figure 5), print quality ratings monotonically decrease with the visually-weighted difference metric. The rank-order correlation between print quality ratings and the visually-weighted difference metric is significant ($p \leq .05$) for both test characters: -0.93 and -0.99 for 'W' and '/', respectively.

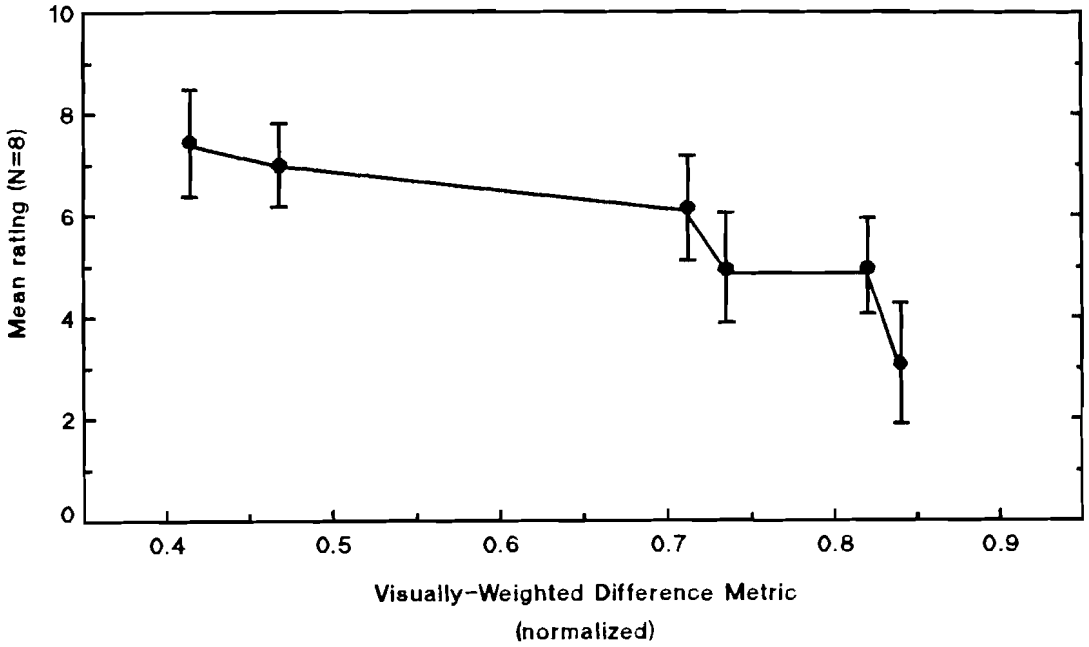

Figure 5. Mean print quality ratings plotted a function of normalized visually-weighted differences between the 'W' character printed with offset lithography and with each condition of inkjet technology.

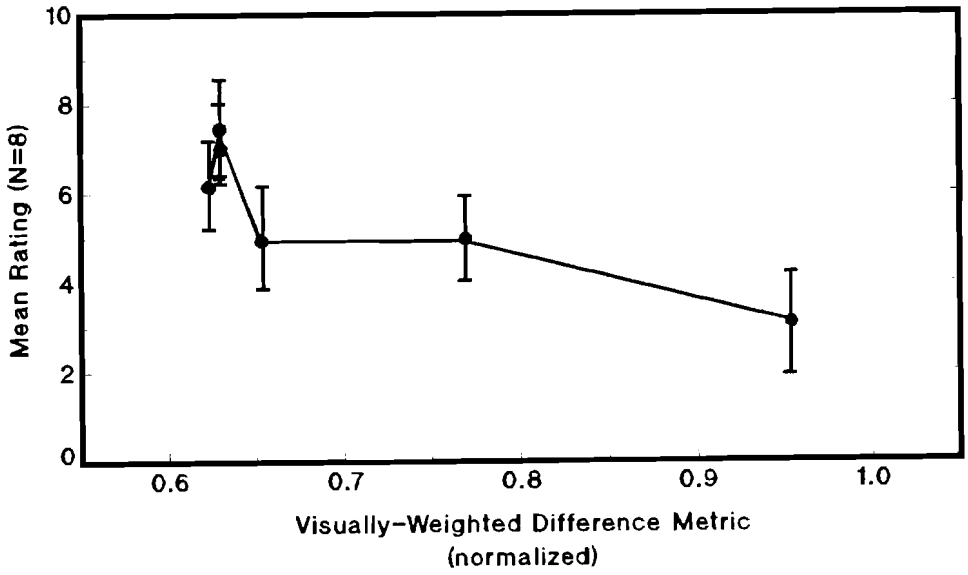

Figure 6. Mean print quality ratings plotted a function of normalized visually-weighted differences between the '/' character printed with offset lithography and with each condition of inkjet technology.

The slight improvement in the relationship between print quality ratings and the visually-weighted difference metric based on the '/' character may reflect the fact that the metric is rotationally variant: Two characters that are the same but differ by a slight rotation will yield different metric values. The fact that the '/' characters were easier to line up visually (the '/' characters were rotated until they appeared to be horizontal lines) may have reduced the amount of rotational error in the camera digitization process. The dependance of the metric on the particular orientation of the letter is one of the limitations of the metric.

Another limitation of the metric is that it does not distinquish between filtered differences of equal magnitude that have different spatial arrangements. Rather, the metric combines all the point by point differences into a single number, thereby losing the spatial information about where the difference originated from. In other words, the metric is phase invariant.

Despite these inherent limitations, the metric does appear to capture some of the general trends in the print quality rating data. For example, Figures 5 and 6 support the assumption that that subjective image quality of printed text is inversely related to the visually-weighted difference between the printed text and the same text rendered with offset lithography. The larger the visually-weighted difference between an inkjet print sample and the higher-resolution offset print sample, the lower the perceived print quality.

Clearly, much can be done to improve the predictions of the metric: For example, Figure 7 shows that although, the visually-weighted difference metric predicts the improvement in print quality with increasing printer resolution and the introduction of grayscale, it does not predict that 150 dpi grayscale and 180 binary have equivalent perceived print quality. One way to improve the relationship between subjective print quality ratings of text and the visually-weighted difference metric is to extend the metric to the analysis of strings of characters or entire paragraphs or pages of text.

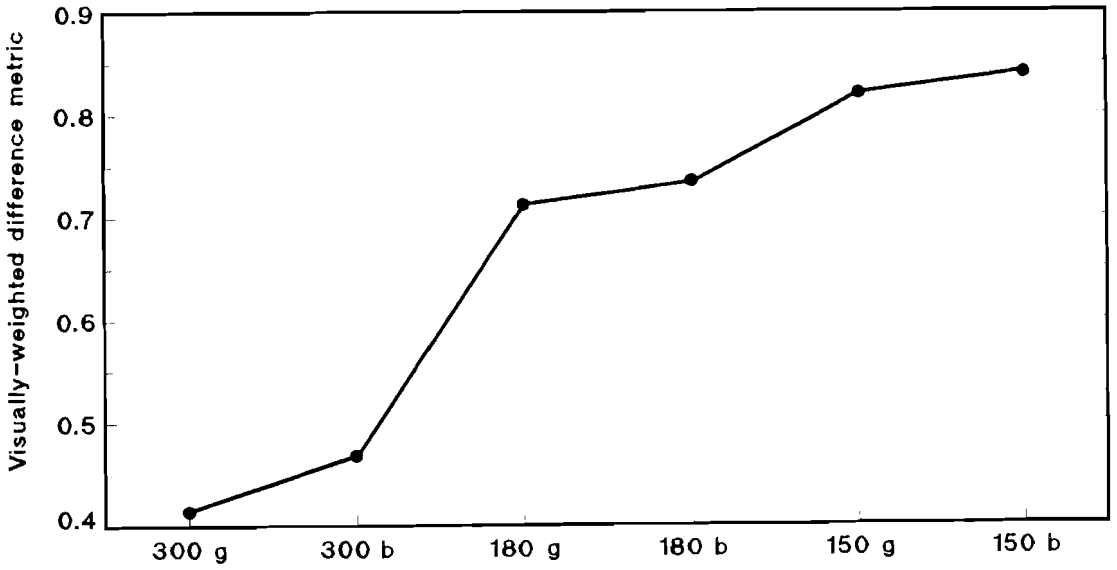

Figure 7. Visually-weighted difference metric plotted for each condition of inkjet technology.

Conclusions

The number of grayscale characters one can create at each sampling resolution is infinitely large. Each type of convolution filter has several parameters to vary. Moreover, shifting the sampling grid with respect to the sampled character (see Figure 1) will result in a different grayscale character. Because it not feasible to empirically evaluate all possible grayscale characters, machine vision metrics become all the more important in the design of high-quality grayscale fonts.

This paper considers two types of machine vision metrics: attribute-based and image-based metrics. Attribute-based metrics require us to make a guess about the features in printed text, business graphics or images that people selectively attend to when judging print quality. Without some guidance from our subjective impressions, we are often at a loss about which print attributes to measure. However, people are often able to rate their subjective impressions of print quality without explicit knowledge of the dimensions or attributes of print quality: People may be able to decide that one print sample looks better than another, but not be able to explain their decision. Even when we have successfully identified print quality attributes, we cannot be confident that subjects will weight the attributes similarly for different print samples. Attributes that are positively correlated with print quality under some conditions, such as edge sharpness, may be negatively correlated with print quality in other conditions. In other words, the way people weight relevant print quality attributes may change depending on the nature of the print samples. Different print samples may cause people to selectively attend to some attributes while ignoring others.

The main advantage of image-based metrics, such as the visually-weighted difference metric, is that it does not assume apriori knowledge about the relationship between subjective print quality and known print quality attributes. Rather, the metric assumes that the smaller the perceived difference between an image of a printed character (or characters) is to an image of the same character(s) printed on an ideal high-resolution printer, the higher the subjective print quality. The assumption that print quality is monotonically related to the perceived similarity between the print sample and an idealized standard is implicit in subjective evaluations of print quality in which people are shown examples of an 'ideal' or standard print sample, such as produced by offset lithography, and asked to rate the quality of print samples relative to that standard. Subjective print quality is then a measure of the perceptual similarity between the rated print samples and the standard print sample. The results of this study suggest that subjective print quality is monotonically related to the difference between a visually-weighted image of a printed test character and a visually-weighted image of the same test character printed with offset lithography.

References and Notes

1. Engeldrum, P. G. (1989) The critical few print quality attributes: What needs to be improved, talk presented at the *8th Annual Print Quality Seminar*, sponsored by BIS CAP International, Boston, Massachusetts, November 3, 1989.

2. Hamerly, J. R. (1981) An analysis of edge raggedness and blur, *Journal of Applied Photographic Engineering, 7:6*, pp. 148-151.

3. Farrell, J. E. & Fitzhugh, A. E. (1989) An image quality metric for digital letterforms, *Topical Meeting on Applied Vision, Technical Digest Series, 16,* (Optical Society of America, Washington, D.C. 1989), pp. 104-107.

4. Judd, D. B & Wyszecki, G. (1975) *Color in Business, Science and Industry*, New York: John Wiley & Sons.

5. Kelly, D. H. (1974) Spatial frequency selectivity in the retina, *Vision Research, 15*, pp. 665-672.

6. Klein, S. A. & Levi, D. M. (1985) Hyperacuity thresholds of 1 sec: theoretical predictions and empirical validation, *Journal of the Optical Society of America, 2:7*, pp. 1170-1190.

7. Campbell, F. W. (1968) The Human eye as an optical filter, *Proceedings of the IEEE, 56:6,* pp.1009-1014

8. Crawford, B. H. (1936) The dependence of pupil size upon external light under static and variable conditions, *Proceedings of the Royal Society, (London), B121*, p. 373.