Quantifying Print Quality for Practice

Elisa H. Barney Smith; Boise State University; 1910 University Dr., Boise, ID, 83725, USA; E-mail: EBarneySmith@BoiseState.edu; Internet: http://coen.boisestate.edu/EBarneySmith
Eric Maggard; Hewlett Packard; 11311 E Chinden Blvd, Boise, ID, 83725, USA; E-mail: eric.maggard@hp.com
Scott Line; Hewlett Packard; 11311 E Chinden Blvd, Boise, ID, 83725, USA; E-mail: kenneth.line@hp.com
Mark Shaw; Hewlett Packard; 11311 E Chinden Blvd, Boise, ID, 83725, USA; E-mail: mark.q.shaw@hp.com

Abstract

To aid in the process of evaluating print quality, five print quality metrics and methods to measure them have been developed. The attributes of interest are: (1) Edge quality, Sharpness, Detail, Raggedness; (2) Scatter, Particles, Halo, Character ghosts; (3) Readability, broken characters (4) Readability, touching characters; (5) Inverse text. The print quality is measured from a test chart containing typical text in a range of sizes and multiple fonts. The test chart is scanned on a commercial desk top scanner. Quantitative values are returned without human input or human surveys, but relate to human perception of these quantities.

Introduction

Printers have become faster and less expensive. Their development cycles have been reduced. Still it is desirable to build products that produce high quality images. Standards exist for print quality [1] and others are under development [4] that look at printed text strictly from a print quality perspective. Spatter of toner particles, edge raggedness and drop out are the types of defects that are of interest here.

Methods have been developed that evaluate the quality of text as a whole to predict a computer's ability to recognize the character symbols given the degradation level and type [2, 3, 5, 9]. These are based at looking at characteristics that were found to confound Optical Character Recognition (OCR) systems in a 1994 study [8]. OCR systems particularly have difficulties with touching and broken characters. They appear often in photocopied documents. The photocopy process is a combination of a scanning process and a printing process. Low quality in either part of the process will result in the image degradation.

This paper builds on these methods to return quantitative values that have a high correlation to perceptual samples. The following attributes are of interest: (1) Edge quality, Sharpness, Detail, Raggedness; (2) Scatter, Particles, Halo, Character ghosts; (3) Readability; (4) Inverse text. Readability was divided into two parts to consider instances of broken characters and touching characters separately, resulting in a total of five quality attributes. The goal is to measure the print quality from a sample test chart containing typical text in a range of sizes and multiple fonts. Quantitative values should be returned without human input or human surveys, but that should relate to human perception of these quantities.

This paper will detail the process used and show the results so the reader can confirm that the procedure has the intended effect. It starts with a description of a perceptual text quality study and the data. Then the methods for extracting the attribute measurements is described. The results are displayed, followed by the conclusion and future work.

Perceptual Study

A perceptual study was conducted to evaluate print quality on the test chart shown in Figure 2a. The test chart was printed on eight printers, that were a mixture of HP and competitive products based on market segments. HP's print quality and image quality experts were asked to evaluate each document for the following text quality attributes:

1. Edge quality - Sharpness - Detail - Raggedness
2. Scatter - Particles - Halo - Character ghosts
3. Text Density - Blackness
4. Hollow characters
5. Readability
6. Inverse text
7. Overall text quality
8. Solid Area Density (SAD) (measured).

They assigned a score for each printer in the test. Each attribute was assessed independently, in isolation, discounting all other attributes or defects, and blind (not knowing the printer which produced the samples). Reviewers were asked to use the following scores while assessing the above attributes:

1. Very Poor/insufficient
2. Lacking
3. Sufficient
4. Good
5. Excellent.

While there was a perceptual difference in the quality for each attribute between printers, the range of values which participants reported was very low.

Figue 1 shows the scores given by the reviewers for the four quality attributes under consideration. They range between 3.375 and 4.625, a response of less than good to better than good. Blurriness had the widest range of score with a difference of 1.125. Readability had a range of only 0.775. This narrow range of responses does not provide a lot of information. Information which engineers can use to evaluate or improve their product is needed.

The original plan was to measure values of attributes that are representative of each quality attribute that correlate with the survey results and generate a mapping that would produce those survey result values from the measurements. Then measurements from an image from a new printer could be used to generate a score that would reflect what the reviewers would have stated.

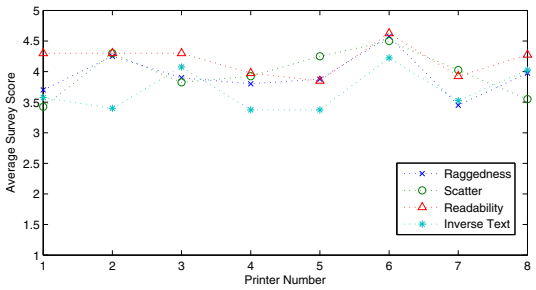


Figure 1. *Survey rating scores given by the print and image quality experts for the 8 printers under evaluation for four quality attributes. The scores show a very narrow range of values.*

Many different measureable attribute quantities were explored. Attribute values were found that had moderate correlation (between ± 0.6 and ± 0.9) with the survey results. This was too low to produce a reliable mapping given the narrow range of survey values, even if the range was spread. Instead the approach was taken to generate values that subjectively to the author team matched with observed degradation levels.

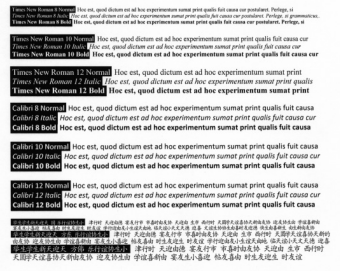
Test Chart

The evaluation is done on a test chart that can be printed by each printer under evaluation. It was designed to print the content that a consumer would likely print and not special test chart elements such as star targets or bar charts. This will allow it also to be perceptually evaluated by humans. At the same time the chart was designed to contain some special marks such as lines, both solid and dotted, which aid automated analysis.

The test chart developed for this work is shown in Figure 2b. It includes text in four fonts. This includes a serif font (Times New Roman), a sans serif font (Calibri), a script font (Coronet), and a font with Chinese characters (DF Ming). The serif and sans serif text appear in both standard and bold. All appear at 4 pt, 6 pt, 10 pt and 12 pt. Straight lines were available, as well as text in each font in inverse video.

Samples of the test chart were produced on eight different printers. The printing was done using multipurpose paper, default print drivers, and Adobe Acrobat software. All the images were scanned at 1200 dpi on a Epson Expression 10000XL scanner with optical resolution of 2400 dpi. Each page was scanned at a preset brightness and contrast to assure consistency between samples. All samples were scanned twice.

Any method can be used to segment the image. Because it is a fixed image pattern at a fixed resolution, much apriori information about the sizes of the various fields can be used. Variations from positioning on the scanner, including both translation and rotation add complexity. We can limit the amount of skew that is present in the images through control of the scanning process, but still some skew is present. We used a variation on the classic X-Y Cut page segmentation method [6] for image segmentation. This problem is well suited to registration to a template image, so long as the template is rotated to the image and not vice versa, as the interpolation necessary for the rotation of the acquired image will introduce undesirable distortions that will affect the measurements.



(a)

(b)

Figure 2. *Samples of test charts used during this work. (a) The test chart used for the perceptual study, (b) the final HP version.*

Quality Attributes

From the initial group of eight quality attributes, four were selected for further development. The attribute "readability" was divided into two attributes, because two different degradations that are common in printing (stroke thickening and stroke thinning) lead to different images affecting readability, and by their separation, the developers can better focus on adjusting the printer performance for the degradation present. The goal was to produce a quantitative measure from a high resolution scan of a page on a commercial desk-top scanner. These numerical values should bear a visual correlation with perceived levels of these degradations. Techniques to measure five quality attributes are described next.

Raggedness

The raggedness quality attribute evaluates edge quality, edge sharpness, edge detail and edge raggedness. All these are related to how the edges of lines vary in position. Raggedness is calculated using a variation of the raggedness attribute from the ISO 13660 standard [1]. Raggedness in the ISO standard is defined as the standard deviation of the edge boundary from the mean of its position. This requires the use of lines or edges.

The (horizontal) underlines of the four font headers were used to measure this attribute. The ISO specification calls for measuring the standard deviation of the distance of the edge to its mean at a threshold level of 60% below the maximum reflectance:

$$R_{60} = R_{max} - 60\%(R_{max} - R_{min}). \quad (1)$$

The maximum and minimum reflectances in the image, R_{max} and R_{min}, are measured. The whole image is then normalized to a $[0,1]$ range. At each column the maximum R_p was found (the darkest point, typically in the middle of the line). Then the first row above or below that which met or fell below the R_{60} threshold was selected. These points were fit to a line with least squares. Points that were more than 2 standard deviations from the average

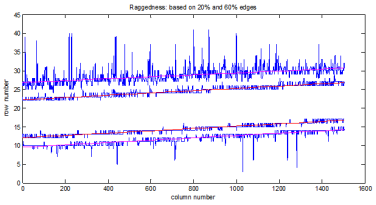


Figure 3. *Example of variation of edge at 20% and 60% reflectance.*

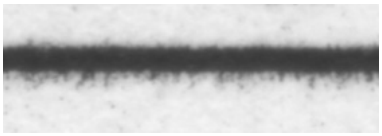


Figure 4. *The amount of raggedness is different for the leading versus the trailing edge relative to the paper feed.*

were excluded as noise to get a better estimate of the line position. Based on this line position estimate, all the edge points were used to calculate the standard deviation of the positions.

Experiments with measurements of the standard deviation of the 60% reflectance showed that for the purpose of developing a raggedness metric, this measurement was too stable. It did not capture enough of the edge variations, and was too similar across print samples with different perceived raggedness. Therefore a second measurement at 20% reflectance was also used. Figure 3 shows the positions of the 20% and 60% reflectance for a printed line. The higher variation in the 20% reflectance edge can be seen. The resulting metric uses a combination of both these reflectances.

The standard deviation of the edge was measured on both the top and the bottom edges of the stroke. The physical movement of the paper relative to the printer results in different effects, Figure 4. The measurements for the 20% reflectance measurements were weighted less than the measurements for the 60% reflectance measurements. The four biased standard deviation measurements were averaged and the measurements were averaged across the four line samples. The raw measurements were then biased so values approximately in a $[0,10]$ range resulted:

$$Raggedness = 0.5 * \sum_{FONT=1..N_{lines}} \frac{(Rag_{20T} + Rag_{20B})/2}{N_{lines}} + \sum_{FONT=1..N_{lines}} \frac{(Rag_{60T} + Rag_{60B})/2}{N_{lines}} \quad (2)$$

This value could be subtracted from 10, if it is desired for 10 to represent how "good" a printer is relative to raggedness, rather than how ragged the printed lines appear.

Scatter

Scatter is also described as particles, halo or character ghosts. The small speckles of toner that are not attached to the character are the elements measured as scatter. The scatter measurement is based on the Small Speckle Factor (SSF) OCR text quality attribute [9]. SSF is a count of all the N_4 connected components (CC) containing fewer pixels than 50% the x-height of the font divided by the number of CCs in the image, N_{CC},

$$SSF = \sum_{n=1..N_{CC}} \frac{CCsize < (0.5 * X\_height)}{N_{CC}}. \quad (3)$$

The connected components are measured from an image thresholded at 157. This was an empirically chosen value, that is close to the Otsu threshold [7] for the image. The x-height is the number of pixels between the baseline and the x-height of a character, Figure 5. Measurements were taken from the regular text lines. The x-height was found by taking a horizontal projection, smoothing it, and looking for the first and last peaks above 75% of the maximum. For the fonts and textual content in this test chart, this method of estimating the text's x-height works well. SSF is normalized by the number of connected components in the input image, N_{CC}.

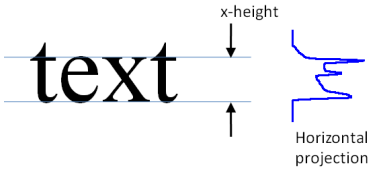


Figure 5. *The measurement of the text's x-height.*

The SSF was calculated for all Latin text lines in the document. The Scatter score is formed from averaging the SSF over all $N_{TextLines} = 48$ regular text lines in the test chart:

$$Scatter = \sum_{n=1..N_{TextLines}} \frac{SSF_n}{N_{TextLines}}. \quad (4)$$

Readability - Broken Characters

Readability can be thought of in many ways. It generally reflects the comfort a reader has reading a piece of text. Because an extended period of time is needed to measure reading comfort, the definition and approach are often modified to instead measure text degradations that affect the recognition of the characters, and thus the words.

Readability is divided into two components: touching characters and broken characters, as both affect readability, and they appear as separate effects from different sources in documents. Touching and broken characters are also attributes in OCR text quality and the methods to measure these quantities are taken from the OCR community [3].

To measure Readability-broken characters, the image is thresholded at a low value to encourage almost broken characters to break. At low thresholds more characters should be broken if they are weak and light in their printing. At high thresholds more characters should be touching if the characters are close to each other and smudged. From the thresholded image, the Broken Character Factor (BCF) is measured from the Times 4 pt text lines.

The N_4 connected components (CC) in the image sample are identified. A tally is kept of which CC heights and widths are present across all CCs in the sample. The BCF calculation revolves around measuring the diversity of the sizes of the connected components in the image. Only CCs with sizes in the BCF "zone," Figure 6, are of interest. The BCF zone is defined to be

the area where the width and height are within 75% of the average connected component height and width, and the height and width differ by no more than 15 pixels. A 2-D histogram is calculated of the CC heights and widths. The number of the size bins that are filled with at least one CC sample within the "BCF zone" are counted. This is normalized by the area of the "BCF zone:"

$$BCF = \frac{number\,of\,width, height\,bins\,filled}{number\,bins\,in\,BCF\,zone}. \quad (5)$$

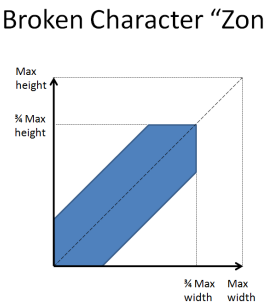


Figure 6. *Values of connected component heights and widths that are within the "BCF Zone."*

The BCF is calculated on an image thresholded at an empirically chosen level of 127. It was applied only to the Times font in 4 pt normal text. This font was thin enough to break. The Coronet font was also prone to breaking and touching characters and is good for human perceptual evaluation of this effect, but its design makes false alarms in the count in this numerical method likely. The sans-serif font is not as prone to breaking, because it maintains a consistent stroke width that is thick enough not to break under normal conditions. There are two 4 pt Times text lines in the testchart. The 2 BCF measurements are summed and scaled to produce the Readability-Broken Characters score:

$$BrokenScore = \sum_{n=1..2} BCF_n * 10. \quad (6)$$

Readability - Touching Characters

Readability - Touching Characters is measured by the Touching Character Factor (TCF). Touching characters are defined by Souza [9] to be characters whose height is less than 3 times the x-height of the character, but contain more than 3 times the x-height number of pixels.

The image is thresholded at an empirically chosen threshold of 190. For each N_4 connected components in the image sample, the height, width and number of pixels is calculated. If the height is less than 3 times the x-height of the character, and the number of pixels is greater than 3 times the x-height number of pixels it is counted as a character instead of a line or a speckle. If a character's height to width ratio is less than 0.75, the TCF ratio value suggested by Souza, the character is considered a touching character. Note that for the fonts present in this test chart, the m's and w's will be considered as touching characters. Since they will be touching in all print samples this will only result in a bias in the score values. The number of touching characters is divided by the total number of characters in the image to form the TCF metric:

$$TCF = \frac{number\,of\,Touching\,Characters}{number\,of\,characters\,in\,text\,sample}. \quad (7)$$

The TCF was calculated for the Times font text samples in 4 pt normal text. The results for each of the two text lines are summed and scaled to produce the score:

$$TouchingScore = \sum_{n=1..2} TCF_n * 10. \quad (8)$$

Inverse Text

The inverse text was used to see how the dot gain affects those text samples. Because the content of the test chart is fixed a priori, this was achieved by measuring how many white pixels appeared in the 4 pt inverse text zones relative to the size of the zone,

$$InvertedCount = \sum_i \sum_j \frac{I(i,j) > \theta}{I(i,j)}. \quad (9)$$

$I(x,y)$ is the inverse text block cropped so the border pixels that are affected by the scanner's point spread function will not contribute to the measurements. The threshold θ = 157 was chosen empirically. The counts from each of the 4 pt inverted text blocks in all fonts are then summed and scaled:

$$InvertedTextScore = \sum_{n=1..6} InvertedCount_n * 15. \quad (10)$$

The value 15 is used to scale this score to be near the [0,10] range. An appropriate baseline, such as an ideal image extracted from a pdf, can be used to scale the values to a desired range.

Results

The estimation of quantitative measures for each quality aspect was applied to the scans of the test chart on eight different printers. 2400 dpi was found to work well, but requires storage and computation that was not necessary, because the metrics were sensitive enough to make useful measurements at 1200 dpi. The scanning resolution of 1200 dpi was used throughout this work. 600 dpi was found to remove image detail that was necessary to measure the metrics, especially raggedness and scatter. Scores were calculated for each of the five quality metrics developed. The scores were sorted and the samples compared to see if the resulting numbers reflected the degradations seen in the images.

The full page images scanned from printing on the eight printers were automatically segmented. Each of the five metrics was calculated using the appropriate parts of the page. Figures 7 - 11 show examples of portions of these pages and the corresponding metric values. In each figure the sub-images are sorted by metric value. In each case the degradation under test can be seen to vary in quantity relative to each metric's numerical value. The fine nuiances in these degradations can be seen, even though the human reviewers rated them all with scores in a small range of values.

Conclusions and Future Work

A practical method to measure five image quality attributes from a test chart that is easy to use has been developed. The returned values relate to the degradations present in the images and can be used to evaluate the quality of the printed sample. The images used in this study are all of relatively high quality, because they are the printers in competition in the marketplace. Evaluation on prints with higher levels of degradation would show the

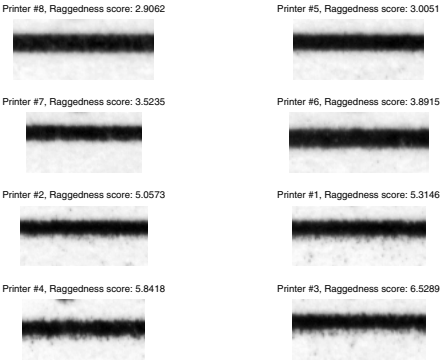


Figure 7. *Samples of strokes illustrating raggedness. The numerical scores are correlated with the perceived level of raggedness.*

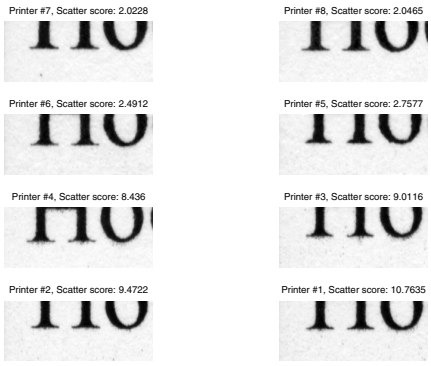


Figure 8. *Samples of strokes illustrating scatter. The numerical scores are correlated with the perceived level of scatter.*

Figure 9. *Samples of strokes illustrating readability-broken characters. The numerical scores are correlated with the perceived level of broken characters.*

Figure 10. *Samples of strokes illustrating readability-touching characters. The numerical scores are correlated with the perceived level of touching characters.*

Figure 11. *Samples of strokes illustrating raggedness. The numerical scores are correlated with the perceived level of raggedness.*

robustness of these metrics. Attempts to map mesurements to the perceptual survey results were not successful because of the narrow range of survey responses.

All the empirical thresholds chosen for this work will need to be reset for any other system. These are based on scanning the test images in gray scale on a particular scanner with particular brightness and contrast settings. All other scanners will have different responses, but similar thresholds for other scanners should not be difficult to find.

Metrics to potentially measure text density, blackness, hollow characters and SAD have been developed, but the samples from the eight printers used in this study did not show enough varaitions in blackness or hollowness to continue to evaluate their effectiveness. The five measurements developed could be combined to form an overall quality score. This requires a decision on the weighting of each attribute.

Experiments were done with test charts (not shown) that included the capital and lower case A-Za-z for each font and size. The word level structure of text, even nonsense, proved to work better, probably because some of the metrics were designed and optimized for use in the OCR domain.

Future work would recommend modifying the test chart slightly to make the segmentation process easier. Greater spacing below the horizontal lines, or if the presence of lines explicitly for that purpose were acceptable, including them in non-interfereing locations woud help.

References

[1] ISO/IEC 13660:2001 Information Technology - Office equipment - Measurement of image quality attributes for hardcopy output - Binary monochrome text and graphic images, 2001.

[2] Luis R. Blando, Junichi Kanai, and Thomas A. Nartker. Prediction of OCR accuracy using simple features. In *Proceedings of the Third International Conference on Document Analysis and Recognition*, pages 319–322, Montreal, Canada, August 1995.

[3] Michael Cannon, Judith Hochberg, and Patrick Kelly. Quality assessment and restoration of typewritten document images. *International Journal on Document Analysis and Recognition*, 2:80–89, 1999.

[4] Edul N. Dalal, Elisa H. Barney Smith, Frans Gaykema, Allan Haley, Kerry Kirk, Don Kozak, Mark Robb, Tim Qian, and Ming-Kai Tse. INCITS W1.1 standards for perceptual evaluation of text and line quality. In *Proceedings SPIE Electronic Imaging, Image Quality and System Performance VI*, volume 7242, page 724203, January 2009.

[5] Juan Gonzalez, Junichi Kanai, and Thomas A. Nartker. Prediction of OCR accuracy using a neural network. In *Proceedings International Workshop on Document Analysis Systems*, pages 323–337, Malvern, PA, 1996.

[6] M.S. Krishnamoorthy, G. Nagy, S. Seth, and M. Viswanathan. Syntactic segmentation and labeling of digitized pages from technical journals. *IEEE Transactions Pattern Analysis and Machine Intelligence*, 15(7):737–747, July 1993.

[7] N. Otsu. A threshold selection method from gray level histograms. *IEEE Trans. Syst. Man Cybern.*, 9:62–66, 1979.

[8] Steve Rice, Frank Jenkins, and Thomas Nartker. *The fourth annual test of OCR accuracy*. UNLV Information Science Research Institute, December 1994.

[9] Andrea Souza, Mohamed Cheriet, Satoshi Naoi, and Ching Y. Suen. Automatic filter selection using image quality assessment. In *Proceedings International Conference on Document Analysis and Recognition*, pages 508–511, Edinborough, Scotland, 2003.

Author Biography

Please submit a brief biographical sketch of no more than 75 words. Include relevant professional and educational information as shown in the example below.

Elisa Barney Smith received her BS in computer science from Rensselaer Polytechnic Institute in Troy, NY (1988), and her MS and PhD in computer and systems engineering also from Rensselaer (1989, 1998). She is currently a professor at Boise State University, Boise ID. Her work has focused on image processing and pattern recognition, primarily applied to image quality in OCR and printing. She also applies image processing and machine learning to solve problems in biomedical imaging, materials engineering and geosciences. She is a senior member of IEEE and SPIE, and an editor for the International Journal of Document Analysis and Recognition (Springer).

Scott Line is a lead monochrome print quality engineer at Hewlett Packard, in the LaserJet Hardware division, Boise, Idaho. Scott has over 12 years' experience in monochrome print quality. Scott received his B.S.M.E. with a minor in computer science from the University of Idaho (1998). Scott received his M.S. in Mechanical engineering with an emphasis in acoustics from the University of Idaho (2002). His focus is in understanding and solving mechanical influences on print quality.

Eric Maggard received his B.S. in Physics from Northwest Nazarene University in 1991 and his M.S. in Computer Science from Walden University (NTU) in 2006. Eric has worked on image analysis metrics and tools in the LaserJet Hardware division for the last 20 years. He has developed algorithms and tools used to analyze printer and scanner image quality throughout the Printer Division. Eric's research interests include object recognition, image classification, and machine learning.

Mark Q. Shaw is a Senior Color and Imaging Architect at Hewlett Packard, in the LaserJet Hardware division, Boise, Idaho. Mark has over 15 years' experience in the Color and Imaging Industry. Mark received his B.S. in Graphic Media Studies from the University of Hertfordshire, his M.S. degree in Color Science from the Munsell Color Science Laboratory, RIT, and is currently working towards a PhD in Electrical and Computer Engineering, at Purdue University in the department of Signal and Image Processing. Mark's research interests include video coding, multi-spectral color reproduction, color modeling, gamut mapping, color management and image understanding.