*Basics for Tone Reproduction in Digital Imaging Systems**

Donald R. Lehmbeck
XACTIV, Fairport, New York, Research Fellow
Xerox Corporation, Webster, New York (retired)
College of Imaging Arts and Sciences, Rochester Institute of Technology, Rochester, New York (Adjunct Faculty - retired)
Imaging Quality Technology Consulting, Penfield, New York- Consultant

Abstract *The basic concepts of reproducing tones with a digital imaging system are reviewed. A characteristic curve approach describing the relationship between input and output tones under a variety of conditions is explored. The most common methods for reproducing tones in digital systems is some form of halftoning and several methods and characteristics are described with illustrations showing pixel level detail. Some examples, constraints and practical limits are discussed at an elementary level. Extensive references are given*

* *Much of this material is taken with permission of the publishers from D. Lehmbeck and J. Urbach, Image Quality for Scanning and Digital Imaging Systems, Chapter 3, pp133-246, in Handbook of Optical and Laser Scanning, 2nd Edition, Ed by G.F. Marshall and G.E. Stutz (CRC Press, Taylor & Francis Group, Boca Raton, FL, 2012)*

Basics of Reproducing Tones

There are four ways in which electronic imaging systems display or print tonal information to the eye or transmit tonal information through asystem:

1. By producing a signal of varying strength at each pixel, using either amplitude or pulse-width modulation.
2. By turning each pixel on or off, i.e. a two-level or binary system;
3. By use of a halftoning approach, which is a special case of binary imaging. Here, the threshold for the white–black decision is varied in some structured way over very small regions of the image, simulating continuous response. Many, often elaborate, methods exist for varying the structure; some involve multiple pixel interactions (such as error diffusion; see the end of Section 3.2.2.3) and others use subpixels (such as high addressability, extensions of the techniques mentioned in Section3.7.2).
4. By hybrid halftoning combining the halftone concept in (3) with the variable gray pixels from (1) (e.g., see References 44 and 45).

From a hardware point of view, the systems are either designed to carry gray information on a pixel-by-pixel basis or to carry binary (two-level) information on a pixel-by-pixel basis. Because a two-level imaging system is not very satisfactory in many applications, some context is added to the information flow in order to obtain pseudo-gray using the halftoning approach.

Macroscopic tone reproduction is the fundamental characteristic used to describe all imaging systems' responses, whether they are analog or digital. For an input scanner or a digital camera it is characterized by a plot of an appropriate, macroscopic output response, as a function of some representation of the input light level. The output may characteristically be volts or digital gray levels for a digital input scanner or camera and intensity or luminance of a display or perhaps darkness or density of the final marks-on-paper image for an output printer. The correct choice of units depends upon the application for which the system response is being described. There are often debates as to whether such response curves should be in units of density or optical intensity, brightness, visual lightness or darkness, luminance, log luminance, gray level, and so on. For purposes of illustration, see Figure 3.13.

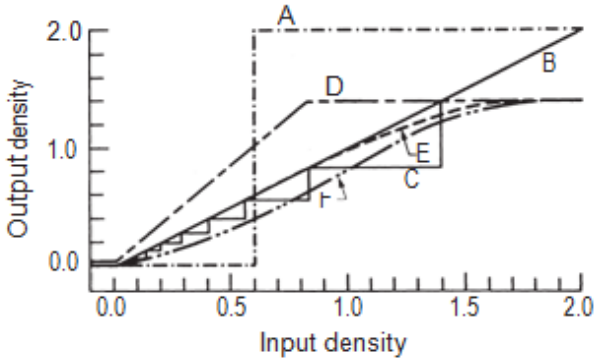


FIGURE 3.13
Some representative input/output density relationships: (A) binary imaging response; (B) linear imaging response; (C) stepwise linear response; (D) saturation–limited linear response; (E) linear response with gradual roll-off to saturation; (F) idealized response curve for best overall acceptability.

Here we have chosen to use the conventional photographic characterization of output density plotted against input density using normalized densities. Curve A (Binary Response) shows the case of a binary imaging system in which the output is white or zero density up to an input density of 0.6, at which point it becomes black or 2.0 output density. Curve B shows what happens when a system responds linearly in a continuous fashion to input density. As the input is equal to the output here, this system would be linear in reflectance, irradiance, or even Munsell value (a system of visual lightness units).

Curve C shows a classic abridged gray (severely limited number of levels) system attempting to write linearly but with only eight levels of gray. This response becomes a series of small steps, but because of the choice of density units, which are logarithmic, the sizes of the steps are very different. Had we plotted output reflectance as a function of input reflectance, the sizes of the steps would have been equal. However, the visual system that usually looks at these tones operates in a more or less logarithmic or power fashion, hence, the density plot is more representative of the visual effect for this image. Had we chosen to quantize in 256 gray levels, each step shown would have been broken down into 32 smaller sub-steps, thereby approximating very closely the continuous curve for B.

When designing a system's tone reproduction, there are many choices available for the proper shape of this curve. The binary curve, as in A, is ideal for the case of reproducing high-contrast information because it allows the minimum and maximum input densities considerable variation without any change to the overall system response.

For reproducing continuous tone pictures, there are many different shapes for the relationship between input and output, two of which are shown in Figure 3.13. If, for example, the input document is relatively low contrast, ranging from 0 to 0.8 density, and the output process is capable of creating higher densities such as 1.4, then the curve represented by D would provide a satisfactory solution for many applications. However, it would create an increase in contrast represented by the increase in the slope of the curve relative to B, where B gives one-for-one tone reproductions at all densities. Curve D is clipped at an input density greater than 0.8. This means that any densities greater than that could not be distinguished and would all print at an output density of 1.4.

In many conventional imaging situations the input density range exceeds that of the output density. The system designer is confronted with the problem of dealing with this mismatch of dynamic ranges. One approach is to make the system respond linearly to density up to the output limit; for example, following curve B up to an output density of 1.4 and then following curve D. This generally produces unsatisfactory results in the shadow regions for the reasons given earlier for curve D. One general rule is to follow the linear response curve in the highlight region and then to roll off gradually to the maximum density in the shadow regions starting perhaps at 0.8 output density point for the nonlinear portion of the curve as shown by curve E. Curve F represents an idealized case approximating a very precisely specified version arrived at by Jorgenson.[56] He found the "S"-shaped curve resembling F to be a psychologically preferred curve among a large number of the curves he tried for lithographic applications. Note that it is lighter in the highlights and has a midtone region where the slope parallels that of the linear response. It then rolls off much as the previous case toward the maximum output density at a point where the input density reaches its upper limit. (See Figure 3.17 later)

Halftone System Response

One of the advantages of digital imaging systems is the ability to completely control the shape of these curves to allow the individual user to find the optimum relationship for a particular photograph in a particular application. This can be achieved through the mechanism of digital halftoning as described below. Historically important studies of tone reproduction, largely for photographic and graphic arts applications, include those of Jones and Nelson,[57] Jones,[58] Bartleson and Breneman,[59] and two excellent review articles, covering many others, by Nelson.[60,61] Many advances in the technology of digital halftoning have been collected by Stoffel[66,67] Sharma[6] (Chapter 6), Loce et. al. [27] and by Eschbach[7]

The halftoning process can be understood by examination of Figure 3.14. In the top of this illustration two types of functions are plotted against distance *x,* which has been marked off into increments one pixel in width. The first type are three uniform reflectance levels, RI, R2, and R3. The second function $T(x)$ is a plot of threshold versus distance, which looks like a series of up and down staircases, that produces the halftone pattern. Any pixels whose reflectance is equal to or above the threshold is turned on, and any that is below the threshold for that pixel is turned off.

Also sketched in Figure 3.14 are the results for the thresholding process for R1 on the second line and then for R2 and R3 on the third line. The last two are indistinguishable for this particular set of thresholding curves. It can be seen from this that the reflectance information is changed into width information. Typically, such threshold patterns (i.e., screens) are laid out two-dimensionally (x and y dimensions) . . An example is shown in Figure 3.15. Therefore the method of halftoning (spatially varying threshold array) is a mechanism for creating dot growth or spatial pulse width modulation over an area of several pixels following the pattern laid out by the thresholding scheme.

This particular thresholding array emulates the printer's 45° screen angle, which is considered to be favorable from a visual standpoint because the 45° screen is less visible (oblique effect[4]) than the same 90° screen (where dots follow the x and y axis) . Other screen angles may also be conveniently generated by a single string of thresholds and a shift factor that varies from raster to raster.[62,63] In Figure 3.15, the numbers in each cell in the matrix represent the threshold required in a 32-gray-level system to turn the system response on or off. The sequence of thresholds is referred to as the dot growth pattern. At the bottom, four thresholded halftone dots (Parts b–e) are shown for illustration. There are a total of 64 pixels in the array but only 32 unique levels. This screen can be represented by 32 values in a 4 × 8 pixel array plus a shift factor of four pixels for the lower set of 32, which enables the 45° screen appearance as illustrated. It may also be represented by 64 values in a single 8 × 8 pixel array, but this would be a 90° screen. It is also possible to alternate the thresholding sequence between the two 4 × 8 arrays, where the growth pattern in each array is most commonly in a spiral pattern, resulting in two unique sets of 32 thresholds for an equivalent of 64 different levels and preserving the screen frequency

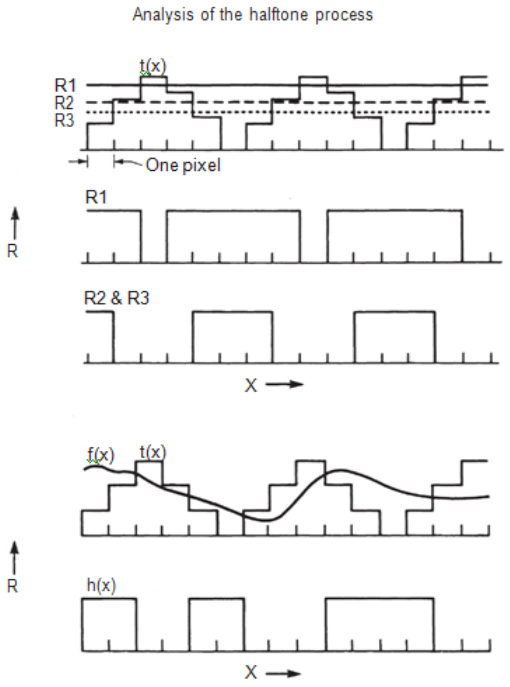


FIGURE 3.14
Illustration of halftoning process. Each graph is a plot of reflectance *R* versus distance *X*. *T(x)* is the proile of one raster of the halftone threshold pattern, where image values above the pattern are turned on (creates black in system shown) by the halftone thresholding process. RI, R2, and R3 represent three uniform images of different average reflectances shown at the top as uniform input and in the middle of the chart as proiles of halftone dots after halftone thresholding. *f(x)* represents an image of varying input reflectance and *t(x)* is a different threshold pattern. *h(x)* is the resulting halftone dot proile, with dots represented, here, as blocks of different width illustrating image variation.

as shown. This screen is called a “double dot.” The concept is sometimes extended to four unique dot growth patterns and hence is named a “quad dot.” Certain percent area coverage dot patterns in these complex multicentered dot structures generate more grey levels but some very visible and often objectionable patterns.

The halftone matrix described in Figure 3.15A represents 32 specific thresholds in a specific layout. There are many alternatives to the size and shape of the matrix, the levels chosen, the spatial sequence in which the thresholds occur, and arrangements of multiple, uniquely different matrices in a grouping called a super cell. Here there are many different cells (more than the four in a quad dot) varying slightly in shape and each may contain a slightly different number of pixels. This gives its designer even more gray levels since there are more cells and each may contain unique thresholds. There are also more available angles due to the size and shape differences of the individual cells giving the centers of the collection of all the supercells more precision to form a new screen angle. See Figure 3.16.

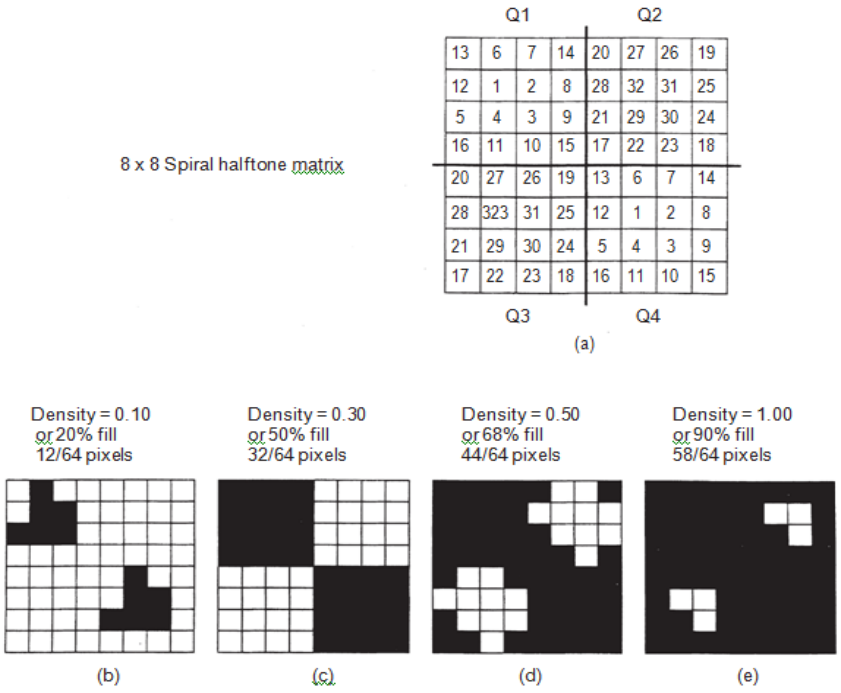


FIGURE 3.15 A
Example of two-dimensional quantized halftone pattern, with illustrations of resulting halftone dots at various density levels.

The careful selection of these factors gives good control over the shape of the apparent tone reproduction curve, granularity, textures, and sharpness in an image. The halftone system’s ability to resolve structures finer than the halftone screen array or cell size has been described as “partial dotting” by Roetling64 and others and is an important and often misunderstood factor in image quality studies (References 53, p. 163; 16, p. 403). It is the result of the high-resolution pixel-by-pixel comparison of the threshold matrix and the image detail which allows high contrast image detail to pass through the halftone matrix, nearly unchanged.

Advanced Concepts and Forms of Halftoning

There are also many other methods for converting binary images into pseudo-gray images using digital halftoning methods of a more complex form.[66,67] These include alternative dot structures, that is, different patterns of sequences in alternating repeat patterns, random halftoning, and techniques known as error diffusion. In his book *Digital Halftoning,* Ulichney[68] describes five general categories of halftoning techniques:

1. Dithering with white noise (including mezzotint)
2. Clustered dot ordered dither
3. Dispersed dot ordered dither (including "Bayer's dither")
4. Ordered dither on asymmetric grids
5. Dithering with blue noise (actually error diffusion)

Figure 3.15 B shows examples of four of these categories. Note the coarseness of the various patterns and the detail rendering differences. Blue noise methods are usually created by error diffusion in which the placement of a single pixel dot is evaluated against the effects of other pixels in its neighborhood and the darkness that is supposed to be rendered there. This difference is an error and is used to determine the placement of the next dot. The dot separation is therefore the variable. If one chooses to look at this spacing as it's reciprocal, that is a spatial frequency. Since it is being changed or modulated, some technologists refer to this as Frequency Modulation. The clustered dot ordered dither halftones have a uniform dot spacing but use dot size i.e. the amplitude of the signal at fixed points to represent the darkness. Some refer to these methods as amplitude modulation.

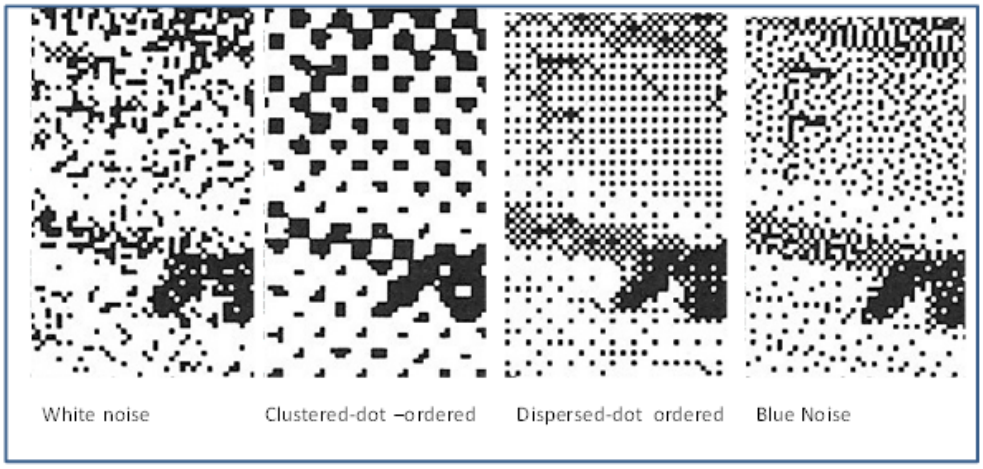


Figure 3.15 B Examples of Ulichney's categories 1,2,3 and 5 all applied to the same part of the same image, Note the variation in the shapes sizes and placement of the dots. Each image shows the same size pixel, easily seen as the tiniest square dot in each case. Each dot pattern here is arranged on a rectangular grid as in Figure 3.15A. Fig 3.15A above is in the clustered dot – ordered dither category

The adjustability of the halftone angles is particularly important in color printing where the interaction of the dot patterns made with different color inks (cyan, magenta, yellow, black) can create disturbing more patterns. Adjusting the supercell precisely to achieve certain specific angles can minimize and even more or less eliminate the effects of these interactions. So in addition to adding many more grey levels, these super-cells offer important color optimization options. Ulichney states that "spatial dithering is another name often given to the concept of digital halftoning. It is perfectly equivalent, and refers to any algorithmic process that creates the illusion of continuous tone images from the judicious arrangement of binary picture elements." The process described in Figures 3.15A and 3.16 falls into the category of a clustered dot ordered dither method (category 2) as a classical rectangular grid on a 45° base.

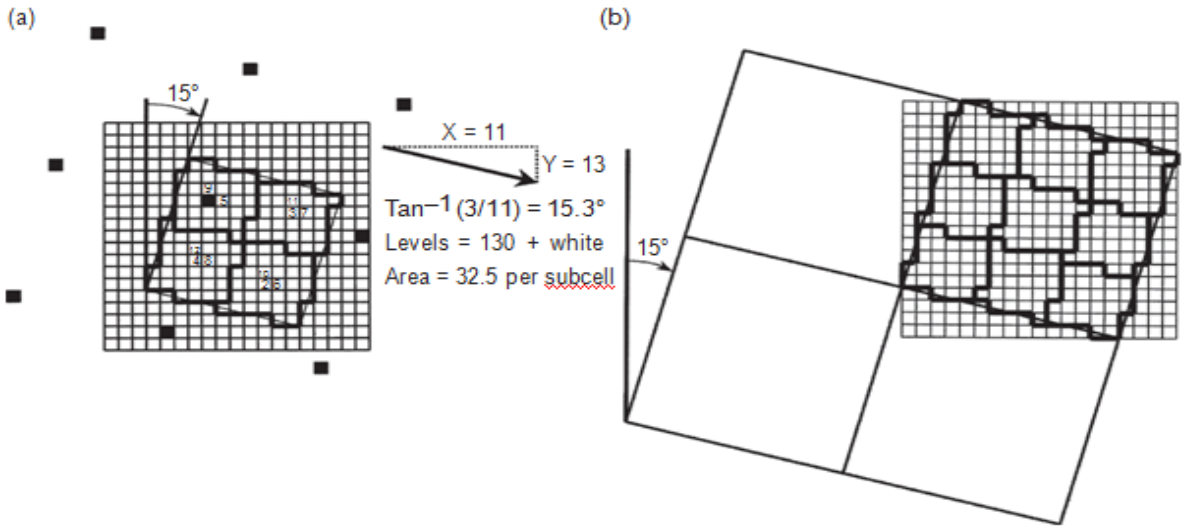
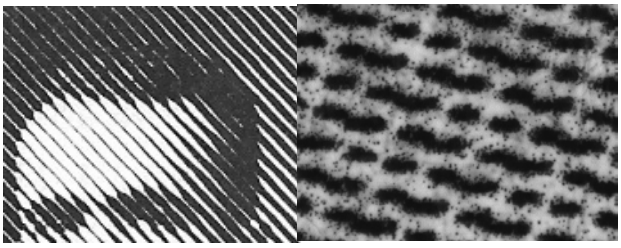


FIGURE 3.16
Examples of multicentered dots: a) a classic quad dot showing the first three thresholds in each individual cell and the large black dots showing the repeat pattern centers at 15.255° and b) a nine center "supercell" where the cell shape and size varies: from L to R 26, 27,27; 27,29,25; 27,27,26 pixels and the angle is 18.4°. (Reproduced with permission from ISO-TC-42, ISO 14524-1999 and 12232:2006(E), International Standards Organization, Geneva, Switzerland, 2006, p. 412 in Chapter 6 by Haines, Wang and Knox.)

Another method of creating a clustered dot, ordered dither pattern is to set up the halftone matrix to cause the dot to grow in a long narrow fashion. This is configured in such a way as to link up the adjoining long dots and making the width of the line represent the darkness of the image in that area. This actually creates a line pattern called a line screen. See Figure 17. For very light areas this would look like isolated dots but they would be arranged along the same lines as the fuller screen at darker levels. Very dark areas would be white dots aligned to the screen direction.



a b

Figure 17 a) A close up of an image made with a with a line screen and **b)** a highly magnified image segment of such a digital line screen showing pixel structure and indicating a multiple dot approach in this case.

There is no universally best technique among these or the many other variations that could not be covered here. Each has its own strengths and weaknesses in different applications. The reader is cautioned that there are many important aspects of the general halftoning process that have not been covered here. (See References 6, 45 and 67 for summaries of digital halftoning technology and many references.)

For example, the densities described in Figure 3.15A only apply to the case of perfect reproduction of the illustrated pixel maps on non-light-scattering material using perfect, totally black inks. In reality, each pattern of pixels must be individually calibrated for any given marking process. The spatial distribution interacts with various noise and blurring characteristics of output systems to render the mathematics of counting pixels to determine precise density relationships highly erroneous under most conditions. This is even true for the use of halftoning in conventional lithographic processes, due to the scattering of light in white paper and the optical interaction of ink and paper. These affect the way the input scanner, a density measuring device or the human eye "sees" a lithographic halftone original. Some of these relationships have been addressed in the literature, both in a correction factor sense[69,70] and in a spatial frequency sense.[71–73] All of these methods involve various ways of calculating the effect that lateral light scattering through the paper has on the light reemerging from the paper between the dots.

The effects of blur from the writing and marking processes involved in generating the halftone, require individual density calibrations for each of the dot patterns and each of the dithering methods that can be used to generate these halftone patterns. The average change in dot size from ideal is called dot gain. The control afforded through the digital halftoning process by the careful selection and calibration of these patterns and methods enables the creation of useful shapes for the tone reproduction curve for a given picture, marking process, or application.

Finally, returning to the type of graph in Fig 3.13 we show a representative characteristic curve in Fig 3.17 for a digital camera(source of the Log H, i.e. Log exposure) vs printer (Source of the D, i.e density) systems which operate with plenty of grey levels but limited by typical optical and materials effects. Note: for readers not familiar with exposure units, H, they are equal to units of illuminance in the image x the time of exposure. For a low flare optical system, Illuminance in the image is proportional to luminance (e.g. brightness) of the scene component that produced it. The middle of the curve is a smooth, typically straight line, the slope of which portrays the contrast of the mid-tones in the image. The shoulder-like shadow region here which describes the gradual loss in slope from the straight portion toward a maximum density in the dark portion of the curve is limited by the maximum density that the printer can produce (around 2.0 for the case shown, glossy photographs) It is often more like 1.4 for typical ink based prints on paper. The darkest exposure regions are often made lighter (more

Log H) by optical flare light added by the camera lens, moving the shoulder to the left.

The toe-like highlight region turns up toward a flat slope limited by the smallest dots that the printer can reproduce, the size of a single pixel at best, often a couple of pixels. A step down from that to pure white paper often occurs and looks like an edge between lightest grey and white paper, this is called a contour. Often the halftone process is arranged to limit the lightest areas printed to the minimum dot size, never showing white base paper. A ***minimum useful output density*** is one which allows the discrimination between two lightest input levels. Similarly a ***maximum useful density*** is the one which allows discrimination between two of the darkest input levels. These are essentially locations of minimum (but not zero) slopes and must be determined for each system and application, The corresponding range of log exposures is a ***useful exposure range*** and is information valuable to setting camera (or other capture device) exposure limits. Landmark studies[1] show the average outdoor scene has a luminance range (~Exposure range) of 160:1(log = 2.2) but can vary up to 700:1(log = 2.8) and down into the middle double digits.

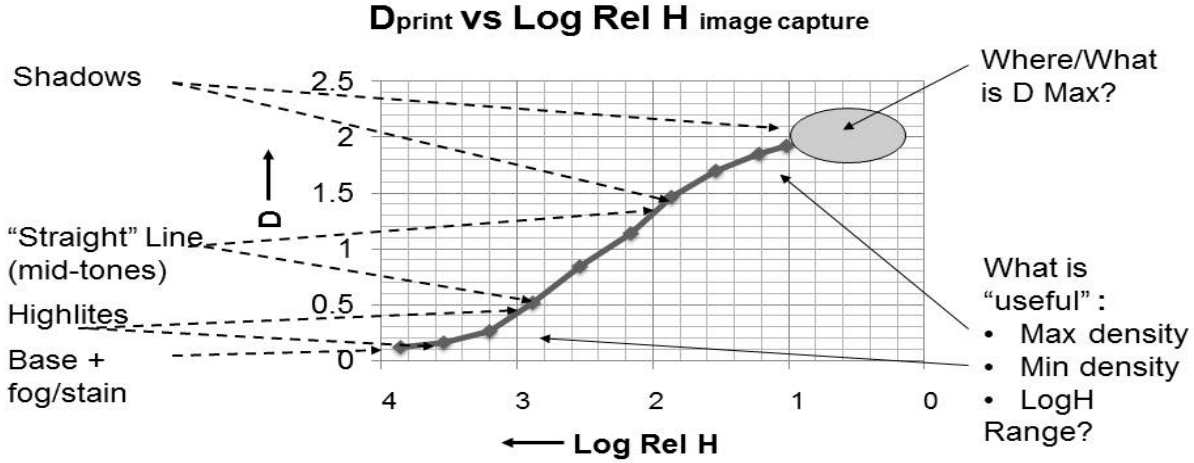


Fig 3.18
The basic representation of tone reproduction for an imaging system is a characteristic curve such as this. The case shown is for a good camera/printer imaging system . It is plotted as an output density shown here for a glossy print vs. an input exposure, shown in. Rel (*relative* to each other, i.e. normalized) Log Exposure units. (note A log of 1= 10 linear units i.e.darker and 4 = 10000 linear units i.e. very bright, i.e. a range of 1- 1000) The commonly referred to parts of the typical tone reproduction curve are indicated on the left. Some useful characteristics to assess from the curve are shown on the right. Various conventions may show either axis plotted in ascending or descending order, different from this version. Among the many variations for such a curve when no print is involved the vertical axis could instead be the log brightness for a display, or log digital output values both of which would flip the y axis giving decreasing values toward the top to keep the curve shape similar

References listed in the order given in the text (and corresponding to the book chapter references**)**

56 Jorgensen, G.W. Preferred tone reproduction for black and white halftones. In *Advances in Printing Science and Technology;* Banks, W.H., Ed.; Pentech Press: London, 1977; 109–142

57 Jones, L.A.; Nelson, C.N. The control of photographic printing by measured characteristics of the negative. *J. Opt. Soc. Amer.* 1942, 32, 558-619

58 Jones, L.A. Recent developments in the theory and practice of tone reproduction. *Photogr. J.* Sect. B 1949, *89B,* 126–151.

59 Bartleson, C.J.; Breneman, E.J. Brightness perception in complex ields. *J. Opt. Soc. Am.* 1967, *57,* 953–957.

60 Nelson, C.N. Tone reproduction. In *The Theory of Photographic Process,* 4th Ed.; James, T.H., Ed.; Macmillan: New York, 1977; 536–560.

61 Nelson, C.N. The reproduction of tone. In *Neblette's Handbook of Photography and Reprography: Materials, Processes and Systems,* 7th Ed.; Sturge, J.M., Ed.; Van Nostrand Reinhold: New York, 1977; 234–246.

66. Stoffel, J.C. Graphical and Binary Image Processing and Applications; Artech House: Norwood, MA, 1982; 285–350.

67. Stoffel, J.C.; Moreland, J.F. A survey of electronic techniques for pictorial image reproduction. IEEE Trans. Comm. 1981, 29, 1898–1925.

6. Hains,C.,Wang,S.,Knox,K., Chapter 6, Digital Color Halftones in Sharma, G. Ed. *Digital Color Imaging Handbook;* CRC Press: Boca Raton, FL, 2003.

7. Eschbach, R. Ed. Recent Progress in Digital Halftoning I and II; Society for Imaging Science & Technology: Springield, VA, 1994, 1999.

27. Loce, R.; Roetling, P.; Lin, Y. Digital halftoning for display and printing of electronic images. In Electronic Imaging Technology; Dougherty, E.R., Ed.; SPIE Press: Bellingham, WA, 1999.

62. Holladay, T.M. An optimum algorithm for halftone generation for displays and hard copies. Proceedings of the SID 1980, 21, 185–192.

63. Roetling, P.G.; Loce, R.P. Digital halftoning. In Digital Image Processing Methods; Dougherty, E.R., Ed.; Marcel Dekker: New York, 1994; 392–395.

64. Roetling, P.G. Analysis of detail and spurious signals in halftone images. J. Appl. Phot. Eng. 1977, 3, 12–17.

53. Lehmbeck,D.R.; Urbach,J.C.; Chapter 3 in Marshall, G editor;. Handbook of Optical and Laser Scanning, chapter 3; Marcell Dekker, NY, 2004 (previous edition of the book & chapter which this article was first printed in).

16. Sharma, G. Digital Color Imaging Handbook; CRC Press, Boca Raton, FL, 2003. An excellent in depth review of many topics including: fundamentals, psychophysics, color management, digital color halftones, compression and camera image processing and more.

65. ISO-TC-42, ISO 14524-1999 and 12232:2006(E), International Standards Organization, Geneva, Switzerland, 2006, See Table 10, this chapter. OECF stands for Opto-electronic Conversion Function—As applied to cameras in ISO 14524 which is conceptually the same for scanners. 12232 deals with speed metrics derived from OECF'

66. Stoffel, J.C. Graphical and Binary Image Processing and Applications; Artech House: Norwood, MA, 1982; 285–350.

67. Stoffel, J.C.; Moreland, J.F. A survey of electronic techniques for pictorial image reproduction. IEEE Trans. Comm. 1981, 29, 1898–1925.

68. Ulichney, R. Digital Halftoning; The MIT Press: Cambridge, MA, 1987.

69. Clapper, R.; Yule, J.A.C The effect of multiple internal reflections on the densities of halftone prints on paper. J. Opt. Soc. Am., 43, 600–603, 1953, as explained in Yule, J.A.C. Principles of Color Reproduction; John Wiley and Sons: New York, 1967; 214.

70. Yule, J.A.C.; Nielson, W.J. The penetration of light into paper and its effect on halftone reproduc- tion. In Research Laboratories Communication No. 416; Kodak Research Laboratories: Rochester, NY, 1951 and in TAGA Proceedings, 1951, 3, 65–76.
71. Lehmbeck, D.R. “Light scattering model for predicting density relationships in reflection images.” Proceedings of 28th Annual Conference of SPSE, Denver, CO, 1975; 155–156.
72. Maltz, M. Light-scattering in xerographic images. J. Appl. Phot. Eng. 1983, 9, 83–89.
73. Kofender, J.L. “The Optical Spread Functions and Noise Characteristics of Selected Paper Substrates Measured in Typical Reflection Optical System Conigurations,” MS thesis, Rochester Institute of Technology: Rochester, NY, 1987.
Jones,L.A.; Condit,H.R., J. Opt. Soc. Amer. 38,123 (1948), 39, 94 (1949) (studied 126 outdoor scenes)