

The State of the Art of Document Image Degradation Modeling

Henry S. Baird

Xerox Palo Alto Research Center
3333 Coyote Hill Road, Palo Alto, CA 94304 USA
baird@parc.xerox.com,
<http://www.parc.xerox.com/istl/members/baird>

Abstract. *The literature on models of document image degradation is reviewed, and open problems are listed. In response to the unpleasant fact that the accuracy of document recognition algorithms falls drastically when image quality degrades even slightly, researchers in the last decade have intensified their study of explicit, quantitative, parameterized models of image defects that occur during printing and scanning. Several models have been proposed, some motivated by the physics of image formation and others by the surface statistics of image distributions. A wide range of techniques for estimating parameters of these models has been explored. These models, in the form of pseudo-random generators of synthetic images, permit, for the first time, investigations into fundamental properties of concrete image recognition problems including the Bayes error of problems and the asymptotic accuracy and domain of competency of classifier technologies. The use of massive sets of synthetic images, in the construction and testing of high-performance classifiers, has accelerated in the last few years. Open problems include the search for methods for comparing competing models and sound methodologies for the use of synthetic data in engineering.*

1 Introduction

This paper reviews the literature on explicit, unambiguous, quantitative models of document-image degradation. Images of paper documents are almost inevitably degraded in the course of printing, photocopying, FAXing, and scanning, and this loss of quality — even when it appears negligible to human eyes — can be responsible for an abrupt decline in accuracy by the current generation of text recognition (OCR) systems. This fragility of OCR systems when confronted by low image quality is well known to OCR engineers and has been illustrated compellingly in experiments carried out at the Information Science Research Institute of the University of Nevada [RKN92] [RNN99].

In addition, there is growing evidence that significant improvement in accuracy on recalcitrant image pattern recognition problems now depends as much on the size and quality of training sets as on choice of image features and classification algorithms. To mention only one example, a U.S. National Institute of

Standards and Technology (NIST) competition on hand-printed digits [Wil92] had a surprising outcome: the competitor with the highest accuracy ignored the training set offered by NIST, using instead its own, much larger, set; furthermore, in spite of widely-divergent algorithms, most of the competitors who used the same training set were tightly clustered in accuracy; and, one of the most promising attacks relied on perhaps the oldest and simplest of algorithms, nearest-neighbor classification [Sab93].

These observations suggest that significantly higher accuracy may be achievable through — and perhaps *only* through — deeper scientific understanding of image quality and the representativeness of image data sets. Such a research program may be expected to assist engineers by allowing them to measure image quality, to control the effects of variation in quality, and perhaps to construct classifiers automatically to meet given accuracy goals.

This survey is organized as follows. First, I describe those degradations that appear to be most important in document image analysis. I summarize the recent history of image quality measurement that is relevant to documents. I describe the degradation models that have been proposed, together with methods for estimating their parameters. I give examples of four types of applications of these models: for the automatic construction of classifiers, in the testing of systems, the provision of public-domain image databases, and in theoretical investigations. Finally, I list open problems.

This integrates and updates much material that appeared separately as [Bai90] [Bai93a] [Bai93b].

2 Document Image Degradations

By “degradations” (or, “defects”) we mean every sort of less-than-ideal properties of real document images, *e.g.* coarsening due to low digitizing resolution, ink/toner drop-outs and smears, thinning and thickening, geometric deformations, etc etc.

These are all departures from an ideal version of the page image which, in the domain of machine-printed textual documents, is usually unambiguously well-defined. In fact we can usually consider such a page’s contents not as an explicit image but as a symbolic representation of an implicit image, in which printing symbols (characters) occur only as references to ideal prototype images in a given library of typeface artwork, together with instructions for their idealized placement (translation, scaling, etc) on the page surface. In practice this idealized symbolic version exists concretely, expressed in PostScript, troff, or similar low-level page description or typesetting languages. But we should not assume that the printing apparatus is always a modern computer-driven typesetting machine. And, in many cases, the symbolic layout description may never have been written down: but it is enough for our purposes that it could have been.

Once the sizes and locations of all symbols and other artwork have been specified, then it makes sense to speak of the ideal image of the page (or text-block, text-line, symbol, etc). Since the document image degradation model literature

focuses almost exclusively on images of high-contrast (essentially monochromatic, black and white) machine-printed pages, I will assume that the idealized input image is bi-level, a two-coloring of the real plane, and thus at an effectively infinite spatial sampling rate.

The defective images resulting from printing and imaging are also commonly bi-level images with a finite spatial sampling rate (often the same, as I will generally assume, along both axes of a conventional X-Y coordinate system, in units of square pixels per inch). The coarsening that results from finite spatial sampling is of course considered to be an image defect.

It's worth noting a few general characteristics of image defects:

- defects are determined by the physics of the apparatus used in printing and imaging; by “apparatus” I mean to include some human actions, as for example when a page is manually placed on a flatbed scanner;
- images may result from more than one stage of printing and imaging; and
- some aspects of the physics are uncontrollable, and so must be modeled as random events and analyzed statistically.

The physics may of course have both ‘global’ and ‘local’ effects. The entire page image is affected by geometric deformations such as non-zero “skew” (rotation). Only a single character image may be affected by ink drop-out. Isolated pixels be individually affected by thermal sensor noise. Thus defects can occur per-page, per-symbol, and per-pixel — and at other levels of the document hierarchy, such as text blocks and text lines. But not all defects can be associated with the symbolic layout: for example, paper grain and ink non-uniformity produce defects larger than a pixel but smaller than a symbol. Thus defects occur over a range of scales, and some — but not all — can be associated with logical parts of the document. These are important considerations in the design of the model, and they are likely to vary with the application.

There is far less agreement, of course, about the details of the physics of printing and imaging. Clearly, there is a great diversity of mechanisms, which include these:

- defocusing;
- binarization (*e.g.* fixed and adaptive thresholding);
- paper positioning variations (skew, translation, etc);
- flaking and poor adhesion generally of ink/toner to paper;
- low print contrast;
- non-uniform illumination;
- pixel sensor sensitivity variations;
- typesetting imperfections (pasted-up layouts, slipped type);
- abrasion, smudging, and creasing of the paper surface;
- spatter, scatter, and spreading of ink/toner;
- vibration and other non-uniform equipment motion;
- noise in electronic components (both additive and multiplicative);
- irregular pixel sensor placement (*e.g.* not lying on a perfectly square grid);
- finite spatial sampling rate;

- non-flat paper surface (*e.g.* curling and warping); and
- non-rectilinear camera positioning (*e.g.* perspective distortion).

Although this list is certainly incomplete, it includes all causes that have been treated quantitatively in the literature to date.

Many images result from multiple stages of printing and imaging. Also, “printing” and “imaging” do not exhaust the possibilities: image compression and transmission (as in FAX) create important classes of defects. Future technologies will no doubt cause defects that we do not now imagine.

3 The Measurement of Image Quality

In this section I will briefly review the state of the recent literature, technical and professional activities, standards, and products, that are relevant to the document image degradation models. I emphasize monochrome and bi-level images, and neglect color.

To the best of my knowledge no comprehensive quantitative study of image defects in printing and imaging has been published. Book-length surveys of a wide range of printing technology, both pre-computer and computer-based, are available (*e.g.* [Phi68], [Sey84], and [McL88]), but image defects are not a principal theme, and are rarely discussed quantitatively. [Mar91] surveys recent optical scanner technology. [Sch86] gives a thorough review of the basic physics of electronic imaging systems, including a helpful chapter on the formation of images of printed documents. In a discussion on “dot size change,” he offers this useful generalization: “many of these stages [of printing and imaging] can be characterized as analog image transfer [involving defocusing] followed by a sharp threshold process.”

Technical meetings that attract papers on this subject include: (a) the annual conference of the Society for Imaging Science and Technology, sponsored by [IST]; (b) the annual Congress on Non-impact Printing Technologies, sponsored by [IST]; (c) the annual Electronic Imaging Science and Technology Symposium, co-sponsored by [IST] and [SPIE]; and of course (d) this meeting. The proceedings of these conferences occasionally contain descriptions of methods for measuring image quality.

The American National Standards Institute [ANSI] subcommittee X3A1, in addition to its work on OCR-A and OCR-B typefaces (X3.17-1981, X3.49-1982, X3.93M-1981, X3.111-1986), has developed standards for ink (X3.86-1980), paper quality (X3.62-1979), and print quality (X3.99-1983).

The barcode industry has been somewhat more attentive to print quality standards than the OCR industry. ANSI subcommittee X3A1, together with committee MH10.8 on Materials Handling, jointly developed a barcode print quality guideline (X3.182-1990). According to its authors [Bis93], it was the first attempt quantitatively to “enable accurate prediction of symbol performance in a reading environment.” It focuses on two-element (bar and space), one-dimensional bar codes. It describes methods for measuring properties of the analog one-dimensional scan reflectance profile such as minimum and maximum

reflectance, symbol contrast, edge determination, and “defects.” “Defects” is defined as the non-uniformity of reflectance within an element relative to overall contrast. The numerical measures can be combined and converted to one of five symbol “grades.” This standard has been rapidly accepted worldwide as the technical basis for most modern barcode designs. Although it is specialized to methods for one-dimensional readout, I recommend it as an excellent guide to the measurement of high-contrast document images. X3A1 is presently working on guidelines for the evaluation of bar code reader performance.

ANSI committee X9B is developing standards on image capture, with special attention to the quality of printed bank checks necessary to permit OCR of fields including handwritten courtesy amounts. In addition, they have inherited the MICR standards from X3A1. The ANSI committee for Graphic Arts Technology Standards [NPES] is working on a standard for the use of densitometers.

The American Society for Testing and Materials [ASTM] has published methods for evaluating business imager quality (committee F5, especially methods F335, F875, F360) and paper quality (committee D6, especially methods F1125, F807, F1351, F1319). These include the measurement of large area density and background for a variety of modern office copiers, and the susceptibility of paper to smudging, creasing, and abrasion (which affect ink and toner scatter and flaking).

The Association for Information and Image Management [AIIM] develops standards, terminology, and tools (including test targets), for evaluating image quality, among other issues. Specially relevant here are standards on recommended practice for quality control of scanners (MS 44), and a tutorial on image resolution (TR 26).

Test target images are widely used to measure document scanner characteristics, notably resolution, thresholding, and contrast ratio [AII,AIIM]. Two of the most commonly used are: (a) the US Air Force 1951 Test Chart (MIL-STD-150-A), for measuring resolving power of optical systems and imaging materials; and (b) the IEEE Facsimile Test Chart. Applied Image, Inc [AII] supplies several test targets specialized to OCR (Applied Scanner test charts 1, 2, and 3). Standardized test target patterns have tended to reflect the lowest common denominator of industrial practice, and do not support complex evaluations such as distinguishing the effects of point-spread function size from the effects of binarization threshold.

A wide variety of densitometers, for measuring point reflectance of paper documents, are commercially available. Aside from these, only a few methods for image characterization have been used widely enough to sustain a market. Specialized devices to measure the “print contrast ratio” have been available for some time (082A [CE], PCM-II [MBC]); this is the quantity $(R_w - R_b)/R_w$, where R_w = the reflectance of a large white spot (typically 0.25 inch radius), and R_b = the reflectance of a small black spot (typically 8 mil radius). The default binarization threshold of document scanners is often specified in units of print contrast ratio, typically as a value in the range [0.5,0.7]. These devices are also able to measure point reflectance, dimensions of symbols, voids, and ink spatter.

More versatile products for evaluating image quality have recently appeared ([RDM], [CE]). They typically consist of a PC, proprietary software, and a specially calibrated monochrome grey-level document scanner, able automatically to measure print contrast, point reflectance, and layout properties such as the accurate placement of the outlines of fields. Specialized software permits testing of OCR-B images of known size for compliance with standards: this involves semi-automatic superimposition of ideal symbol prototypes on the image, and estimation of symbol height and width and stroke thickness. At present, these products do not support the estimation of point-spread function, thresholding, or pixel sensitivity variations; neither do they analyze test target images.

Test methods for evaluating paper quality (opacity, specular gloss, etc.) are available from [TAPPI].

It is common practice in industrial laboratories to develop application-dependent methods for measuring and controlling image defects; these methods and their associated devices, software, and test target images are often abandoned at the project's end ([Tre93], [Loc93], [Blo93], [Gil93]). Thus, although many systematic quantitative studies of defects in printing and imaging equipment have been carried out, most have been narrowly specialized, and few have been published.

4 Document Image Degradation Models

We introduce the discussion of document image degradation models by a brief account of the central technical issues governing their design. We then briefly describe two models, one based on the physics of printing and imaging, and the other based on surface statistics of image distributions. For both of these we summarize methods that have been proposed for estimating their free parameters in order to fit the models to real image populations.

4.1 Methodological Issues

The central technical questions to ask about a proposed model are these:

- *Parameterization.* Is the model expressible as an explicit computable function of a small, fixed number of numerical parameters? (If not, then it is hard to see how it can be used effectively to solve engineering problems.)
- *Randomization.* Which of the model's effects are intrinsically random? Can their distributions be parameterized (as above)? If so, we include the parameters of their distributions among the parameters of the model.
- *Validation.* For any given defective image for which an ideal prototype is known, what is the probability that there exists some values of the model parameters that, applied to the ideal prototype, will duplicate the defective image? Since realistic models are likely to be probabilistic, the answer to this question must be probabilistic also.
- *Parameter Estimation.* For any given population of defective images with known ideal prototypes, can a distribution on the model parameters be inferred that closely fits the real distribution?

We can distinguish two generic approaches to specifying models. The first is to model the physics of the apparatus in detail. The completeness of such models can then be justified in part by pointing to the physics. Certainly this can lead to accurate models, but they may be unnecessarily specific and complicated. The second approach is more empirical: propose the simplest model that merely “saves the appearances,” that is, that is able to generate duplicates of real defective images. Such models cannot be justified by appeals to physics, and must rest on purely statistical measures of completeness. Models of both types have been proposed.

4.2 A Physics-Based Model

A single-stage parametric model of per-symbol and per-pixel defects, modeled on the physics of printing and imaging, was proposed in [Bai90] and refined in [Bai93a]. The model parameters include:

- **size**: the nominal text size of the output (units of points);
- **resn**: “resolution,” the output spatial sampling rate (pixels/inch);
- **skew**: rotation (degrees);
- **xscl**, **yscl**: multiplicative scaling factors (horizontally and vertically);
- **xoff**, **yoff**: translation offsets (output pixels);
- **jitt**: jitter, the distribution of per-pixel discrepancies of the pixel sensor centers from an ideal square grid: vector offsets (x,y) are chosen for each pixel, each component independently (the standard error of a normal distribution with zero mean, in units of output pixels);
- **blur**: defocusing, modeled as a Gaussian point-spread function (psf) centered at the pixel sensor center (the standard error of the psf kernel in units of output pixels);
- **sens**: sensitivity, the distribution of per-pixel additive noise (the standard error of a normal distribution with zero mean, in units of intensity); and
- **thrs**: the binarization threshold (in units of intensity, where 0.0 represents white and 1.0 black).

When the model is simulated, the parameters take effect in the order given above: the ideal input image is first rotated, scaled, and translated; then the output resolution and per-pixel jitter determine the centers of each pixel sensor; for each pixel sensor the blurring kernel is applied, giving an analog intensity value to which per-pixel sensitivity noise is added; finally, each pixel’s intensity is thresholded, giving the output image. In practice the values of these parameters are chosen pseudo-randomly, for each symbol, from first-order parametric distributions determined by mean and variance values specified by the user.

The input to the pseudo-random generator is an “ideal” black and white image at high resolution: in practice, scalable outline descriptions purchased from typeface manufacturers are used. The pseudo-random number generator is an implementation of the mathematics in [Zei69], whose seed, during long runs, is occasionally reset with low-order bits of the CPU fine-grain timer.

Note that, for all the parameters, new values are chosen randomly for each symbol — thus I call them per-symbol parameters. But, two of these values — pixel sensor sensitivity and jitter — are themselves per-pixel parameters controlling randomization of each pixel. Thus, the values of the per-symbol parameters are subject to direct control (by specifying a constant distribution), but the values of the per-pixel parameters are subject to only indirect control. This has important consequences for Monte Carlo parameter-estimation.

4.3 A Statistics-Based Model

[KHP93] proposes a model of document imaging that includes both global (perspective and non-linear illumination) and local (speckle, blur, jitter, and threshold) effects. The order of application of the model is similar to that described above. The optical distortion process is modeled morphologically rather than by appeal to the physics of blurring and binarization.

4.4 Estimation of Model Parameters

All of the proposed models possess numerous free parameters which must be chosen to fit them to real image populations. Where the printing and scanning apparatus is itself available to be tested, special image patterns ("test targets") can be useful. In other circumstances, we must rely on images not specially designed for the purpose — such as images of machine-printed text in known typefaces — to drive the estimation process.

4.5 Estimation using Images of Test Targets

There appears to be no commercially available test targets for the estimation of these model parameters:

- affine deformations (skew, shear, magnification, etc);
- size of point-spread function kernel;
- threshold; and
- pixel sensitivity variations.

In [Bai93a] a conceptual design of test targets, including sector star targets (*e.g.* [AII], target MT-17), for these parameters was given.

Elisa H. Barney Smith has carried out a thorough analysis [Smi98] of the effectiveness of sector star and other test targets in estimating the operating parameters of scanner systems. Using models of the scanning process and specially designed bi-level test targets, four methods of estimating parameters were developed. One of these estimates the displacement of a scanned edge and the other three estimate the scanner's point spread function (PSF) width and binarization threshold by analysis of particular features of the images of the test targets. These methods were tested systematically using both real and synthetic images. The resulting estimates are close to those implied by analysis of grey-level images. The parameters that were estimated were used to generate synthetic character images that "in most cases bear a strong resemblance" to real images acquired on the same equipment.

4.6 Estimation using Images of Text

If we do not have access to the printing and imaging apparatus, we must attempt to estimate the parameters by a computation on a set of images, some of whose properties may be known.

If the documents are Manhattan textual layouts, it is well known then both skew and shear can often be measured to a small fraction of a degree, by a variety of algorithms (*cf.* [Bai87]).

If the typeface and text size are known, then it may be possible to estimate many of the remaining parameters. [Bai93a] describes some preliminary experiments with *black-box* parameter-estimation, in which the defect model is treated as a black box that can be affected only by varying the model parameters and observing the output defective image.

Attempts to exactly match a given target image with pseudorandomly generated images proved to be futile due to the vast number of distinct images generated. By averaging over a set of “good” (but not identical) matches, estimation succeeds within a reasonable amount of computation, when only one parameter is varied at a time. More realistically, *all* parameters were then treated as unknown during estimation: large-scale runs (requiring 1.8 CPU hours for each set of parameters) were successful in estimating **skew**, **xscl**, and **yscl** accurately and repeatably; the other parameters were less tractable.

Kanungo’s Method Tapas Kanungo [Kan96] proposed a statistical bootstrapping method for rejecting the hypothesis that two image sets (say, one real, the other synthetic) were drawn from the same unknown distribution. In the context of parameterized degradation models, the rate at which Kanungo’s method rejects the hypothesis can be analyzed as a function of a model parameter, so providing a technique for estimating model parameters: experiments have shown ([KHBSM94], [KHB95a], [KHB95b]) it to work reliably and efficiently in the estimation of some (not yet all) parameters of both the [Bai90] and [KHP93] models. In addition it seems to have promise as a method for comparing two competing models: exploration of this remains an important open issue.

In [Bai99], the slightest changes in document image quality that can be distinguished reliably and fully automatically by Kanungo’s method were measured. For six parameters of the [Bai90] model, remarkably fine discriminations are possible, often subtler than are evident to visual inspection. And, as few as 25 reference images are sufficient for this purpose. These results suggest that Kanungo’s method is sufficiently sensitive to a wide range of physics-based image degradations to serve as an engineering foundation for many image-quality estimation and OCR engineering purposes.

5 Applications of Models

Applications of these models have taken many forms. Perhaps their widest impact is in the generation of synthetic data sets used in training classifiers for doc-

ument image recognition systems. They have also been used to carry out large-scale systematic tests of systems. Public-domain software and image databases have been published. Finally, they have permitted a new class of experimental studies of image recognition systems using very large-scale simulation. The following sub-sections give some representative examples of such applications. A comprehensive survey of applications has not yet been attempted, here or elsewhere.

5.1 Constructing Classifiers

In an early experiment [Bai90], a Tibetan OCR system was constructed using training data that was initialized with real images but augmented by synthetic variations. Perhaps the first large-scale use of these models to help construct an industrial-strength classifier was [BF91], in which a full-ASCII, 100-typeface classifier was built using synthetic data only (and tested, with good results, on real images). In a series of similar trials, synthetic data was has been used [MB97] to construct pre-classification decision trees with bounded error (which also worked well in practice). [HB93] discusses an application of image defect generators in the automatic construction of “perfect metrics” (distance functions from an image to a class of images), for use in classifiers exhibiting both high accuracy and excellent reject behavior.

5.2 Testing OCR Systems

[Jen93] describes experiments with synthesized images of complete pages of text, using a model of near-ideal printing and imaging, in support of an effort to measure baseline performance of commercial OCR page readers.

6 Public-Domain Software and Image Databases

Software implementing the model of [KHP93] has been made available as part of the “English Document Database CD-ROM” [PCHH93] designed by The Intelligent Systems Laboratory of the Department of Electrical Engineering, University of Washington, Seattle, WA.

Image defect models and their associated generators permit a new kind of standard image database which is explicitly parameterized, alleviating some drawbacks of existing databases. The first publicly-available database of this kind, the “Bell Labs image defect model database, version 0,” was designed for publication in the CD-ROM mentioned above. The database contains 8,565,750 bi-level images, each labeled with ground truth. The images are of isolated machine-printed characters distorted pseudo-randomly using the image defect model of [Bai90]. It is designed to assist research into a variety of topics, including: (a) measurement of classifier performance; (b) characterization of document image quality; and (c) construction of high-performance classifiers. The ground

truth of each image specifies which symbol it is and its typeface, type size, image defect model parameters, and true baseline location. Each model parameter ranges over a small set of values, and the cross-product of these ranges has been exhaustively generated, to permit the design of systematically fair experiments operating on a wide variety of subsets of the database.

No more than a third of the images are “easy”: that is, only slightly or moderately distorted, and so readily recognizable by most commercial OCR machines. A large number — perhaps a fifth — are “impossible”: that is, distorted so extremely that they can not be recognized by even the best modern experimental OCR algorithms. The rest of the images are distributed, by small steps in parameter space, across the interesting boundary separating easy from impossible.

A speed-up in classification time of nearly a factor of five, achieved without special hardware, is significant and practically important; and in some applications an extra error of at most 0.5% is acceptable. Encouraged by this, we plan to build trees using larger pseudo-randomly generated training sets.

6.1 Simulation Studies

Large-scale simulations [HB97] using generative models permit empirical exploration of important open questions concerning realistic image pattern recognition problems. Consider a given defect model, applied to a given ideal prototype image, as a stochastic source of an indefinitely long sequence of defective images. This induces a probability distribution on the space of all discrete bi-level images. It should be clear that many questions of practical interest about the performance of classifiers can be stated as quantitative properties of these distributions. Unfortunately, in most cases of practical importance it is not feasible, with our present analytical methods and computer algorithms, to describe these distributions explicitly. The difficulties do not all arise from the complexities of defect models: many are grounded in the arbitrary nature of the prototype images, which are analytically-intractable artifacts of human history and culture.

One interesting question is whether or not the Bayes error of a given problem is non-zero, and whether it can be estimated within tight bounds. This question, posed for a two-class discrimination problem (‘e’/‘c’ in FAX-quality images), has been answered: it is now computationally feasible to estimate the Bayes error of such concrete, realistic image recognition problems, within tight bounds.

Another question is whether or not, given an indefinitely large volume of training data, dissimilar classification methodologies will achieve the same accuracy asymptotically. In an experiment on three different trainable classification technologies, all of them capable of increasing their ‘capacity’ (VC-dimension) indefinitely, it has been shown that all three asymptotically approach a classification accuracies that are statistically indistinguishable from one another. Extensions of these experiments to a wider range of classification technologies is an urgent open issue.

Also, it is natural to wonder under what circumstances is classification accuracy a smooth monotonic function of degradation parameters. For the physics-based model of [Bai90] and for classifiers over the ASCII symbol set, it has been

shown that this is the case for most of the parameters. Thus it is possible to map a "domain of competency" of the classifier in degradation parameter space, and furthermore it has been shown in small scale experiments that this domain may be used to select training regimens that improve accuracy at the margins of acceptable performance.

7 Open Problems

Many unsolved problems, both theoretical and practical, remain.

We are approaching a day when researchers and engineers can choose from among several realistic, carefully validated mathematical models of image defects, together with software implementations in the form of pseudo-random defect generators. In particular, although we have made considerable progress in the last decade, we still feel a need for

- a theoretical framework for validating models that provides a rigorous foundation for objective, empirical, and computable criteria for demonstrating the completeness of models, and for comparing competing models;
- algorithms for estimating distributions on all of their model parameters to fit real image populations closely; and
- public-domain, portable, and fast model-simulation software capable of generating images at the character, word, text-line, and page level.

We should also extend our methods to cope with grey-level and color document images.

Further progress on these open problems will, I believe, prove to be critical to progress on a broad array of problems arising in theoretical studies and engineering practice.

7.1 Uses of Synthetic Data

Synthetic data is increasingly being used, along with or instead of real data, in training and testing of recognition systems. This practice has provoked a debate among engineers, which was reflected in a panel discussion [DP99] organized last September at the 5th ICDAR. Six panel members, including this author, spoke to the question "under what circumstances is it advisable to use synthetic data?"

Here is my summary of the points made:

- The classifier that is trained on the most data wins.
- It doesn't matter how much real data you train on: it's never enough.
- Training only on real data feels safe — after all, almost everyone does it — but it isn't.
- Real data is corrupting: it is so expensive that we reuse it, repeatedly, with unprincipled abandon.
- Synthetic data is selfless: it is born only to be used once and then thrown away.

- Everyone should feel free to train on any data he/she chooses, real or synthetic or both.
- Training only on synthetic data is courageous, today — and perhaps foolhardy — but one day it will be wise.
- Training on a mixture of real and synthetic data may be, today, the safest — but we don't know the best proportions.
- Training on mislabeled data is asking for trouble. Training on correctly labeled data, but of low image quality, is just as dangerous.
- We don't know how to separate helpful from unhelpful training data, whether real or synthetic.
- It might help to generate synthetic training data that complements real data by a process of interpolation between real samples.
- It might hurt to generate synthetic training data by extrapolating from real data.
- Testing on synthetic data to claim good performance is unprincipled.
- Testing on synthetic data to identify weaknesses is virtuous.
- Some models are more equal than others, but it's still a mystery which is which.
- Truth is stranger than fiction.
- The document image degradation model research area will live forever, since we will never agree on the models.

Acknowledgements

Much of this work is joint with Tin Kam Ho and Tapas Kanungo. I am indebted to George Nagy, Bob Haralick, Elisa H. Barney Smith, David Ittner, and Theo Pavlidis for stimulating discussions on these subjects. I wish to thank Larry Spitz for providing references [Mal83], [Edi87], and [MS88]; he and Perry Stoll helped me reimplement my model at PARC. I have also benefited from the advice of David Albrecht, Roland Aubey, Chuck Biss, Robert Bloss, Michael Bruno, Ken Church, Allan Gilligan, Robert Gruber, Fred Higgins, Brian Johannesson, Mary Kastner, Robert Lettenberger, Bob Loce, Roger Morton, Del Oddy, Pat Pavlik, Gil Porter, Bud Rightler, Frank Romano, Paul Ross, Timothy Tredwell, and Luc Vincent. Any omissions or misstatements are, of course, mine.

References

- [AII] Applied Image Inc, 1653 East Main St, Rochester, NY 14609.
- [AIIM] Association for Information and Image Management, 1100 Wayne Avenue, Silver Spring, MD 20910. (Formerly the National Micrographics Association).
- [AIM] AIM USA, 634 Alpha Drive, Pittsburgh, PA 15238-2802. (Trade association for automatic identification and keyless data entry technologies.)
- [ANSI] American National Standards Institute, 11 W 42 St, New York City, NY 10036.

- [ASTM] American Society for Testing and Materials, 1916 Race Street, Philadelphia PA 19103.
- [Bai87] H. S. Baird, "The Skew Angle of Printed Documents," *Proc., 1987 Conf. of the Society of Photographic Scientists and Engineers*, Rochester, New York, May 20–21, 1987.
- [Bai88] H. S. Baird, "Feature Identification for Hybrid Structural/Statistical Pattern Classification," *Computer Vision, Graphics, and Image Processing*, Vol. 42, No. 3, pp. 318–333, June 1988.
- [Bai90] H. S. Baird, "Document Image Defect Models," *Proc., IAPR Workshop on Syntactic and Structural Pattern Recognition*, in Murray Hill, NJ, 13–15 June, 1990. Reprinted in H. S. Baird, H. Bunke, and K. Yamamoto (Eds.), *Structured Document Image Analysis*, Springer-Verlag: New York, pp. 546–556, 1992.
- [Bai93a] H. S. Baird, "Calibration of Document Image Defect Models," *Proc., 2nd Annual Symposium on Document Analysis and Information Retrieval*, Caesar's Palace Hotel, Las Vegas, Nevada, pp. 1–16, April 26–28, 1993.
- [Bai93b] H. S. Baird, "Document Image Defect Models and Their Uses," *Proc., 2nd Int'l Conf. on Document Analysis and Recognition*, Tsukuba Science City, Japan, pp. 62–67, October 20–22, 1993.
- [Bai99] H. S. Baird, "Document Image Quality: Making Fine Discriminations," *Proc., IAPR 1999 Int'l Conf. on Document Analysis and Recognition*, Bangalore, India, September 20–22, 1999.
- [BF91] H. S. Baird and R. Fossey, "A 100-Font Classifier," *Proceedings, IAPR 1st ICDAR*, St.-Malo, France, 30 September – 2 October, 1991.
- [Bis93] Charles E. Biss, PSC, Inc., 770 Basket Road, P.O. Box 448, Webster, NY 14580–0448. (Chair, ANSI X3A1.3 Working Group on Image Quality.)
- [Blo93] Robert Bloss, personal communication, UNISYS, 41100 Plymouth Rd, Plymouth, Michigan 48170, March 1993.
- [Bun87] W. Buntine, "Learning Classification Trees," *Statistics and Computing*, vol. 2, pp. 63–73, 1992.
- [CE] Clearwave Electronics, 8701 Buffalo Avenue, Niagara Falls, NY 14304.
- [CN84] R. G. Casey and G. Nagy, "Decision Tree Design Using a Probabilistic Model," *IEEE Trans. Information Theory*, Vol. IT-30, No. 1, pp. 94–99, Jan. 1984.
- [DP99] S. Dennis and I. Phillips, "Ground Truthing: Real or Synthetic Data – a Panel Discussion," at 5th Int'l Conf. on Document Analysis and Recognition, Bangalore, India, September, 20–22, 1999.
- [Edi87] J. R. Edinger, Jr., "The Image Analyzer — A Tool for the Evaluation of Electrophotographic Text Quality," *Journal of Imaging Science*, Vol. 31, No. 4, pp. 177–183, July/Aug. 1987.
- [Gil93] Allan Gilligan, personal communication, AT&T Bell Laboratories, West Long Branch, NJ, March 1993.
- [HB93] T. K. Ho and H. S. Baird, "Perfect Metrics," *Proceedings, IAPR 2nd ICDAR*, Tsukuba, Japan, October 20–22, 1993.
- [HB97] T. K. Ho and H. S. Baird, "Large-Scale Simulation Studies in Image Pattern Recognition," *IEEE Trans. on PAMI*, Vol. 19, No. 10, pp. 1067–1079, October 1997.
- [IEE92] *Proceedings of the IEEE*, Special Issue on OCR, July, 1992.
- [IST] Society for Imaging Science and Technology, 7003 Kilworth Lane, Springfield, VA 22151.

- [Jen93] F. Jenkins, *The Use of Synthesized Images to Evaluate the Performance of OCR Devices and Algorithms*, Master's Thesis, University of Nevada, Las Vegas, August, 1993.
- [KHP93] T. Kanungo, R. M. Haralick, and I. Phillips, "Global and Local Document Degradation Models," *Proceedings, IAPR 2nd ICDA*R, Tsukuba, Japan, October 20-22, 1993.
- [KHBSM94] T. Kanungo, R. M. Haralick, H. S. Baird, W. Stuetzle, & D. Madigan, "Document Degradation Models: Parameter Estimation and Model Validation," *Proc., Int'l Workshop on Machine Vision Applications*, Kawasaki, Japan, December 13-15, 1994.
- [KHB95a] T. Kanungo, R. M. Haralick, H. S. Baird, "Validation and Estimation of Document Degradation Models," *Proc., 4th Annual Symp. on Document Analysis and Information Retrieval*, Las Vegas, Nevada, pp. 217-225, April 24-26, 1995.
- [KHB95b] T. Kanungo, R. M. Haralick, & H. S. Baird, "Power Functions and Their Use in Selecting Distance Functions for Document Degradation Model Validation," *Proc., IAPR 3rd Int'l Conf. on Document Analysis & Recognition*, Montreal, Canada, August 14-16, 1995.
- [Kan96] T. Kanungo, *Document Degradation Models and Methodology for Degradation Model Validation*, Ph.D. Dissertation, Dept. EE, Univ. Washington, March 1996 [Supervisor: Prof. R. M. Haralick].
- [Knu96] D. E. Knuth, *Computer Modern Typefaces*, Addison Wesley, Reading, Massachusetts, 1986.
- [Loc93] Robert Loce, personal communication, Xerox Webster Research Center, 800 Phillips Road, Webster, NY 14580, March, 1993.
- [Mal83] M. Maltz, "Light Scattering in Xerographic Images," *Journal of Applied Photographic Engineering*, Vol. 9, No. 3, pp. 83-89, June 1983.
- [Mar91] G. F. Marshall (Ed.), *Optical Scanning*, Marcel Dekker: New York, 1991.
- [MB97] C. L. Mallows and H. S. Baird, "The Evolution of a Problem," Special issue of *Statistica Sinica* in honor of H. Robbins, Vol. 7, No. 1, pp. 211-220, January 1997.
- [MBC] Macbeth Corp, P.O. Box 230, Newburgh, NY 12551-0230.
- [McL88] R. McLean, *The Thames and Hudson Manual of Typography*, Thames and Hudson, London, 1988.
- [MS88] M. Maltz and J. Szczepanik, "MTF Analysis of Xerographic Development and Transfer," *Journal of Imaging Science*, Vol. 32, No. 1, pp. 11-15, Jan./Feb. 1988.
- [NPES] National Printing Equipment and Supply Association, 1899 Preston White Drive, Reston, VA 22091.
- [PCHH93] I. T. Phillip, S. Chen, J. Ha, and R. M. Haralick, "English Document Database Design and Implementation Methodology," *Proceedings, 2nd Annual Symposium on Document Analysis and Information Retrieval*, Caesar's Palace Hotel, Las Vegas, Nevada, April 26-28, 1993, pp. 65-104.
- [Phi68] A. Phillips, *Computer Peripherals and Typesetting*, Her Majesty's Stationery Office, London, 1968.
- [Por93] Gil Porter, personal communication, Xerox Webster Research Center, 800 Phillips Road, Webster, NY 14580, March 1993.
- [RNN99] S. V. Rice, G. Nagy, and T. A. Nartker, "OCR: An Illustrated Guide to the Frontier," Kluwer Academic Publishers, 1999.

- [RKN92] S. V. Rice, J. Kanai, and T. A. Nartker, "A Report on the Accuracy of OCR Devices," ISRI Technical Report TR-92-02, Univ. Nevada Las Vegas, Las Vegas, Nevada, 1992.
- [RDM] RDM Corp, 608 Weber St N., Waterloo, Ontario N2V 1K4, Canada.
- [Sab93] M. Sabourin, A. Mitiche, D. Thomas, and G. Nagy, "Hand-Printed Digit Recognition using Nearest Neighbour Classifiers," *Proceedings, 2nd Annual Symposium on Document Analysis and Information Retrieval*, Caesar's Palace Hotel, Las Vegas, Nevada, pp. 397-409, April 26-28, 1993.
- [Sch86] W. F. Schreiber, *Fundamentals of Electronic Imaging Systems*, Springer-Verlag: Berlin, 1986.
- [Sey84] J. W. Seybold, *The World of Digital Typesetting*, Seybold Publications, P.O. Box 644, Media, PA 19063, 1984.
- [Smi98] E. H. Barney Smith, "Optical Scanner Characterization Methods using Bilevel Scans," Ph.D. Dissertation, Computer and Systems Engineering Dept, Rennselaer Polytechnic Institute, December 1998 [Supervisor: Prof. G. Nagy].
- [SPIE] Society of Photo-Optical Instrumentation Engineers, 1000 20th St, Bellingham, Washington, 98225.
- [TAPPI] TAPPI, 15 Technology Parkway South, Norcross, GA 30092.
- [THP93] T. Kanungo, R. M. Haralick, and I. Phillips, "Global and Local Document Degradation Models," Submitted to IAPR 2nd Int'l Conf. on Document Analysis and Recognition, Tsukuba, Japan, 1993.
- [Tre93] Timothy Tredwell, personal communication, Head, Imaging Electronics Lab, Eastman Kodak Research Labs, Rochester, NY, March, 1993.
- [Wil92] R. A. Wilkenson, et al, "The First Census Optical Character Recognition Systems Conference," NIST Internal Report, Gaithersburg, Maryland, 1992.
- [WS87] Q. R. Wang and C. Y. Suen, "Large Tree Classifier with Heuristic Search and Global Training," *IEEE Trans. PAMI*, **PAMI-9**, No. 1, pp. 91-102, Jan. 1987.
- [Zei69] N. Zeirler, "Primitive Trinomials Whose Degree is a Mersenne Exponent," *Inf. Control*, **15**, 1969.