Table 1: Experimental results for the automatic construction of decision–trees for preclassification of 100 typefaces of the full printable ASCII symbol set. 'Fraction' is the fraction of training data used to "grow" the tree; the remainder is used to "populate" it. 'Space' is the maximum main memory process size in Mbytes. 'Time' is the CPU time to build the tree (hours:minutes). 'Depth' is the depth of the tree, averaged over all training samples. 'Speed–up' is the average pruning factor measured during test. The remaining entries are preclassification errors (in %): '%error' is averaged over the five types sizes 6, 8, 10, 12, and 14 point.

fraction	training (1,079,413 samples)			testing (393,030 samples)						
	space	time	depth	speed–up	%error	6p	8p	10p	12p	14p
1/8	5.16	9:23	6.7	4.7	1.0	2.1	1.4	0.8	0.4	0.4
1/7	5.66	10:24	7.2	4.8	1.1	2.2	1.5	0.8	0.5	0.5
1/6	6.30	13:23	7.6	5.2	1.2	2.4	1.7	0.9	0.6	0.5
1/5	7.32	16:53	8.3	5.5	1.5	2.8	2.0	1.1	0.7	0.6
1/4	8.55	21:23	9.1	6.0	1.7	3.3	2.3	1.3	0.8	0.7
1/3	10.46	29:49	10.4	6.8	2.0	3.9	2.6	1.5	1.0	0.8
1/2	*excessive*									

[Knu96] D. E. Knuth, *Computer Modern Typefaces*, Addison Wesley, Reading, Massachusetts, 1986.

[Mal83] M. Maltz, "Light Scattering in Xerographic Images," *Journal of Applied Photographic Engineering*, Vol. 9, No. 3, June 1983, pp. 83–89.

[MS88] M. Maltz and J. Szczepanik, "MTF Analysis of Xerographic Development and Transfer," *Journal of Imaging Science*, Vol. 32, No. 1, Jan./Feb. 1988, pp. 11–15.

[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.

[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.

[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, April 26-28, 1993, pp. 397-409.

[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, Jan. 1987, pp. 91–102.

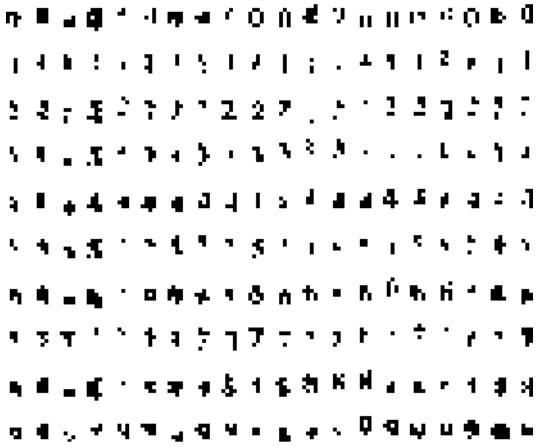

Figure 3: Twenty images each of the ten Times Roman digits 0–9 (top to bottom), generated pseudo–randomly at a type size of 5 point and a spatial sampling rate of 100 pixels/inch. We have shown that, in spite of the extreme distortions, this problem has an intrinsic error rate of only 35%.

observed only 7499 distinct images, so that each distinct image occurred more than six times on average.

We specify further that the classes are *a priori* equally likely, and that the position of images with respect to the origin of the image plane is immaterial. Then no classifier can improve on one which (a) matches an unknown image to a prototype only if, under some translation, they are pixel–for–pixel identical, and then (b) assigns to the image the class that, in the limit, occurs most frequently for that prototype (breaking ties at random). Since such a classifier is optimal for the problem, we call its error the *intrinsic error* of the problem. This can be measured in a brute force fashion as follows: for each distinct image (under translation), make a prototype and label it with the class whose image instances match it most often: then count as errors all image instances whose true class differs from their prototype's label.

On one sample of 1000 images per class ("1000×10"), the observed intrinsic error was 35.1%. On another, distinct, 1000×10 set, we observed 33.8%. On a 5000×10 set, a superset of these, we observed 35.7%. The consistency of these results suggests that the data sets are large enough to be representative.

We compared to this the performance of a nearest-neighbor classifier using Euclidean distance with 1000×10 prototypes (breaking ties at random): on a distinct 1000×10 set, its error was 52.3%.

We have not carried out a controlled psychophysical trial to measure the accuracy of human readers on this problem. Nevertheless, we do not expect that any human subject could achieve close to either of these results.

Although this is a small trial run at an extremely coarse spatial sampling rate, we feel that the results are surprising enough to be potentially controversial. If one grants merely that the image defect model that we used is fairly realistic — it doesn't have to be perfect to make the point — then we have exhibited *a natural text recognition problem on which machine classifiers can significantly outperform humans in accuracy.* We speculate that this gap in capability will hold at larger, more ordinary, spatial sampling rates.

This experiment was feasible, within the time and space constraints of our computing environment, because the number of distinct images that occurred was manageable. Scaling up to larger images will be difficult, but it may be feasible: the number of distinct images that occur, though large, is much smaller than we initially expected, and may not grow exponentially with image size, so that hashing may suffice; and, the generation of images is easily parallelized.

7 Open Problems

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

However, many obstacles, both theoretical and practical, remain. I offer a list of open problems.

- There is an urgent need for further discussion of protocols for validating models of this kind.
- There appear to be serious methodological problems in establishing the completeness of models of realistic complexity.

Acknowledgements

Much of this work is joint with Tin Kam Ho. I am indebted to David Ittner and George Nagy for stimulating discussions on these subjects. I wish to thank Larry Spitz for providing references [Mal83], [Edi87], and [MS88]. My ideas on the validation of models have been clarified in discussions with Ken Church.

References

[Bai92] H. S. Baird, "Document Image Defect Models," in H. S. Baird, H. Bunke, and K. Yamamoto (Eds.), *Structured Document Image Analysis*, Springer–Verlag: New York, 1992, pp. 546-556.

[Bai93] H. S. Baird, "Calibration of Document Image Defect Models," *Proceedings, 2nd Annual Symposium on Document Analysis and Information Retrieval*, Caesar's Palace Hotel, Las Vegas, Nevada, April 26-28, 1993.

[BF91] H. S. Baird and R. Fossey, "A 100-Font Classifier," *IAPR 1st ICDAR*, St.-Malo, France, 30 September - 2 October, 1991.

[Bun87] W. Buntine, "Learning Classification Trees," *Statistics and Computing*, vol. 2, 1992, pp. 63–73.

[CN84] R. G. Casey and G. Nagy, "Decision Tree Design Using a Probabilistic Model," *IEEE TRans. Information Theory*, Vol. IT-30, No. 1, Jan. 1984, pp. 94–99.

[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, July/Aug. 1987, pp. 177-183.

[HB93] T. K. Ho and H. S. Baird, "Perfect Metrics," *Proceedings, IAPR 2nd ICDAR*, Tsukuba, Japan, October 20-22, 1993.

[IEE92] *Proceedings of the IEEE*, Special Issue on OCR, July, 1992.

[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 ICDAR*, Tsukuba, Japan, October 20-22, 1993.

context is an effort to construct a classifier for 100 typefaces of the full ASCII alphabet (details of the protocol are given in [BF91]). In the present trial, 1,472,443 images altogether were generated, and split into two distinct sets: 1,079,413 samples were used for training, and 393,030 for testing.

The decision trees were intended for use as *preclassifiers*: that is, their purpose was to prune the set of classes to a small fraction of their default number, so that more accurate classifiers, executed downstream, would have less to do (their runtime being approximately linear in the number of classes to be distinguished). At each node of a tree, a single binary–valued feature was tested. Each leaf owns a (sub)set of the classes. An input vector of binary features is said to be correctly preclassified if the leaf it determines contains its true class.

The tree–growing heuristic was to pick "splits" (a leaf and feature pair) until the tree acheived a pruning factor of 16. The split chosen was the one which, among all possible next splits, maximally increased the entropy of the tree as a whole. This pruning factor and entropy were estimated on the assumption that the distribution of the training data was representative. By construction, all the trees perfectly preclassify the training data.

Testing reveals, in general, non–zero preclassification error and a pruning factor different (but in the event very close to) the one estimated during training.

Even with the heuristic's short–cuts, problems of this large size push the limits of modern computing technology. We ran on a Silicon Graphics Computer Systems Power Series Model 4D/480S, with 8 40MHZ IP7 processors, running time–shared UNIX. In order for the program to terminate in less than a CPU month, it was necessary hold all training data in main memory during the tree–growing phase. Swapping contention effectively limited the size of a CPU–bound process to 15 Mbytes.

For these reasons, we partitioned the training data into two subsets (at random). The first subset was used to grow the tree, and so was held in main memory. The second was reserved on disk, and, after the tree was constructed, read in one by one merely to update the leaves. Table 2 summarizes the results. In the table, 'fraction' is the fraction of the training data used grow the tree.

Using 1/8 to grow the tree (and 7/8 to populate), we constructed a tree with a speed–up (pruning factor) of $\times 4.7$ and a preclassification error of 1.0% averaged over five type sizes. The errors were concentrated, unsurprisingly, in the smaller type sizes: at 10 and 12 point, the error was less than 0.5%. As we varied the fraction used to grow the tree, we measured a range of trade–offs between speed–up and error, finally observing a speed-up of $\times 6.8$ with an added error of 2.0%.

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 experiment with even larger pseudo–randomly generated training sets.

6 Estimating Intrinsic Error

The ability precisely to describe classes of realistic imaging defects permits progress to be made towards answering fundamental theorectical questions that have been long neglected. I will describe one of these here: for reasons of space, I will suppress the mathematical formalities.

Consider a given defect model, applied to a given ideal prototype image, as a stochastic source of an indefinitely long sequence of defective images. Associated with it, therefore, is a probability distribution on the space of all discrete bi–level images. It should be clear that most 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, essentially non–analytical artifacts of human history and culture.

Still, given pseudo-random image generators, as computable approximations to the ideal stochastic sources, we can conceive of experiments to attack certain problems. For example, it is natural to ask, of a recognition problem, what is its *intrinsic error*: that is, what is the lowest probability of error that the best possible classifier could achieve? Consider the problem of distinguishing **c** from **e**, under a given defect model (say [Bai92]), with fixed size and variable spatial sampling rate. Where the sampling rate is very high, the images are large and only slightly affected by the defect model, and then the question is answerable: the intrinsic error is effectively zero. As the sampling rate decreases, experience tells us, the intrinsic error will become significantly greater than zero: this occurs because the class distributions now overlap. Above a certain threshold, all the generated images vanish, and intrinsic error becomes 1.

These are all gendanken experiments; we can go a step further, and carry out an experiment. We have found a problem for which it is computationally feasible to estimate the intrinsic error. In this problem, there are ten classes, whose ideal prototype images are the ten numeric digits from the Adobe Times Roman typeface. The image defect model is as described in [Bai92], with the spatial sampling rate set to 100 pixels/inch and type size set to 5 point. Note that the problem is not unusual, save for one feature: the spatial sampling rate is extremely coarse. The resulting images are small and, by ordinary standards, highly distorted (Figure 3).

During generation, any image whose bounding box was wider or higher than 5 pixels was discarded (an insignificant fraction); 'null' (vanishing) images were not discarded; finally, five thousand images were kept for each class. Each image was encoded in a translation-invariant manner (as a string), and the occurrences of each distinct image counted.

Note that $2^{5\times 5} \approx 33.6$ M distinct images could potentially have been generated and kept. However, we

- *Completeness.* 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.

- *Calibration.* 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?

Of course, these questions, as stated, are still vague about crucial details. For example, how does one estimate the probability that a particular defective image will be duplicated, when the model itself is partially randomized? There is an urgent need for further discussion of methods for validating models of this kind.

4 A Public-Domain Image Database

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. This section gives a brief description of the first publicly-available database of this kind, the "Bell Labs image defect model database, version 0." This was designed for publication in the "English Document Database CD–ROM" funded by ARPA and designed by The Intelligent Systems Laboratory of the Department of Electrical Engineering, University of Washington, Seattle, WA[PCHH93]. At the time of writing, this CD–ROM is scheduled for publication in the Fall of 1993.

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 [Bai92]. 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.

The alphabet is the full printable ASCII set of 94 symbols. One typeface is represented: Computer Modern Roman, of the Metafont family[Knu96], the artwork for which is in the public domain and is free and widely available in computer–legible form. The spatial sampling rate is fixed at 300 pixels/inch. Five type sizes are represented: 12, 10, 8, 6, and 4 point. Figure 2 illustrates some of the realistic distortions produced by the interaction between the size of the blurring kernel and the binarization threshold.

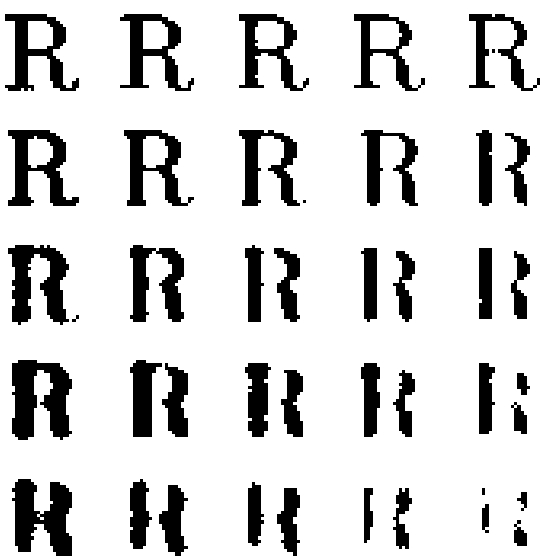

Figure 2: Twenty–five images of the Computer Modern Roman 'R', showing the cross–product of five values each of two parameters: the digitization threshold (increasing left to right), and the size of the blurring kernel (increasing top to bottom).

5 Constructing Decision Trees

Non–backtracking decision trees are an attractive technology for classification, since they promise a dramatic trade–off of time for space. A theoretical drawback is that inferring optimal trees, under various criteria, appears to be computationally infeasible[Bun87]; however, suboptimal heuristics often build roughly balanced, strongly pruning trees[CN84, WS87]. Most such heuristics are greedy: given an incomplete tree, they choose the next split (of a leaf) that is locally most promising: for example, it may, among all possible single next splits, maximize the increase in entropy of the tree as a whole. A serious practical drawback of this strategy is a rapid accumulation of error as the tree deepens, for a training set of fixed size.

It is nevertheless often possible to construct shallow trees with acceptably low error; it is the deeper, strongly pruning — and thus faster — trees that exhibit unacceptable error. This experience suggests that the essential problem is not that the heuristic is sub-optimal: it is that the training data is too sparse. This observation motivated a series of experiments in using image defect generators in a brute force way to reduce the error rate of fast decision–trees.

One such experiment was designed as follows. The

affected by noise. Thus there are defects that occur per–page, per–symbol, and per–pixel, and at other levels of the document hierarchy.

Figure 1: Thirty-six images of the Japanese word 'kekkan' (defect), from the JTeX Mincho typeface, illustrating the range of image defects generated pseudo–randomly by a model discussed in the text. There are twelve samples for each of three point sizes (10, 8, and 6), at a spatial sampling rate of 300 pixels/inch. The images are magnified in order to make the spatial sampling rate visible.

2 The Recent Literature

The measurement of image defects has long been of interest to the optics, imaging, and scanning communities, and has been addressed by standards organizations including ANSI, ASTM, AIIM, and TAPPI: a survey is given in [Bai93]; additional useful references include [Mal83], [Edi87], and [MS88].

[Bai92] discusses motivations for developing defect models that are specific to document images, proposes a research agenda for this purpose, and gives particulars of a ten-parameter model that approximates some aspects of machine-printing and imaging of text, including symbol size, spatial sampling rate (digitizing resolution), affine spatial deformations, jitter, speckle, blurring, and thresholding. Figure 2 illustrates the range of effects generated pseudo–randomly by this model. The discussion is confined to local (per-character and per-pixel) effects. It reports several applications of the model in the construction of custom classifiers with a minimum of manual effort.

[Bai93] reviews the state of the art, including the research literature, standards, professional activities, and commercially available tools such as hardware/software calibration devices and test targets. A refinement of the model of [Bai92] is described, together with experiments in Monte Carlo calibration, with the goal of estimating the model parameters that best fit a given population of images of known typeface. The results suggest that most of its model parameters can be estimated independently of one another, using averages of trials weighted by translation-invariant Hamming distance. Also, a conceptual design of an advanced calibration test target is discussed.

[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.

In this proceedings, [KHP93] discusses a model of document imaging that includes both global (perspective and non-linear illumination) and local (speckle, blur, jitter, and threshold) effects. The optical distortion process is modeled morphologically, and a method for inferring the model parameters is discussed.

Also in this proceedings, [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.

At the time of writing, at least five U.S. research teams are investigating image defect models.

3 Models and Their Validation

We can distinguish two generic approaches to specifying models: *explanatory* and *descriptive*. An explanatory model is based on details of the physics. Such models can be validated in part by pointing to the physics. This can lead to accurate models, in the limit, but they may be unnecessarily specific and complicated. A descriptive model is more empirical, merely "saving the appearances" by closely fitting the data. Such models are validated principally by statistical measures, for example, of the probability of generating duplicates of real defective images. A review of the recent history of modern standards in this area (cf. [Bai93]) suggests that proposals for explanatory models dominate the early stages of debate, but descriptive models eventually gain the consensus necessary for adoption.

The essential technical questions to ask about a proposed descriptive model of document image defects are, I believe:

- *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 must be randomized? Can their distributions be parameterized (as above)? If so, we include the parameters of their distributions among the parameters of the model.

Document Image Defect Models and Their Uses

Henry S. Baird
AT&T Bell Laboratories
600 Mountain Avenue, Room 2C-322
Murray Hill, NJ 07974-0636 USA

Abstract

The accuracy of today's document recognition algorithms falls abruptly when image quality degrades even slightly. In an effort to surmount this barrier, researchers have in recent years intensified their study of explicit, quantitative, parameterized models of the image defects that occur during printing and scanning. I review the recent literature and discuss the form these models might take. I give a preview of a large public–domain database of character images, labeled with ground–truth including all defect model parameters, the first of its kind. I describe the use of massive pseudo-randomly generated training sets for the construction of high-performance decision trees for preclassification. Also, I report preliminary results along a more theoretical line of attack: the estimation of the intrinsic error rate of precise–specified text recognition problems (this is joint work with Tin K. Ho). Finally, I list some open problems.

1 Introduction

In recent years, some researchers in document recognition have voiced the concern[IEE92] that existing methods for designing high-performance classifiers have hit a barrier. This point-of-view is supported by two observations:

1. **The accuracy of today's document recognition algorithms falls abruptly when image quality degrades even slightly.** "Slightly," of course, in the eyes of human readers. This has long been the conventional wisdom among researchers. Recently, it has been illustrated compellingly by experiments [RKN92] carried out at the Information Science Research Institute of the University of Nevada.

2. **Significant improvement in accuracy on hard problems now depends as much, or more, on the size and quality of training sets as on algorithms and hardware.** This opinion, although strongly held by some researchers, is not yet widely accepted. It seems to have been corroborated by the surprising outcome of a recent U.S. National Institute of Standards and Technology (NIST) competition on handprinted digits [Wil92]. The competition's one clear winner ignored the training set offered by NIST, and used instead its own, much larger, proprietary training set. Furthermore, in spite of their widely-divergent algorithms and hardware, most of the competitors who used the same training set were tightly clustered in accuracy. One of the most promising attacks on this problem relies on one of the oldest and simplest of algorithms: nearest-neighbor classification[Sab93].

These observations suggest that research on image quality and the representativeness of image data sets should now be assigned a higher priority than in the past. One issue inextricably involved with both of these topics is image degradation. An empirical scientist, faced with phenomena that he or she hopes to understand better, naturally proposes explicit, unambiguous, quantitative models of it, and attempts to validate the models by fitting them to real data. Such a research program may be expected to assist engineers eventually, 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.

By "defects"[1] I mean a wide variety of less–than–ideal properties of real images. In this paper, I constrain the discussion to models of defects due to the physics of apparatus for printing and imaging (in "apparatus" we must include the people operating the machines). I do not discuss here the exciting nascent literature on models of shape deformations due to parameterized typefaces or individual handwriting style variations.

The physical causes of image defects are myriad: spreading and flaking of ink/toner; paper surface defects; optical and mechanical deformations and vibration; low print contrast; non–uniform illumination; defocusing; finite spatial sampling rate; variations in pixel sensor sensitivity and placement; noise in electronic components; binarization (*e.g.* fixed and adaptive thresholding); and miscellaneous trauma (*e.g.* coffee stains). And, of course images may result from more than one stage of printing and imaging.

The physics may of course include both 'global' and 'local' effects. Spatial deformations may affect an entire page image, while ink dropouts affect only a single character image, and pixels may be individually

[1] The term "degradation" is a reasonably apt alternative, but I prefer "defect" since it is shorter. "Distortion" is also widely used, but it carries the principal sense of "continuous deformation," and as a result sounds awkward — to my ear — when applied to discrete phenomena such as dropouts.