12

The State of the Art of Document Image Degradation Modelling

Henry S. Baird

12.1 Introduction

This chapter reviews the literature and the scientific and engineering state of the art of models of document-image degradation. Images of paper documents are almost inevitably degraded in the course of printing, photocopying, faxing, scanning, and the like. This loss of quality, even when it appears negligible to human eyes, can cause an abrupt drop in accuracy by the present generation of text recognition (OCR) systems. This fragility of OCR systems due to low-image quality is well known by serious users as well as OCR engineers and has been illustrated compellingly in large-scale experiments carried out at the Information Science Research Institute of the University of Nevada ([55] through [53]). 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 representativeness of training sets as on choice of features and classification algorithms. To mention only one example, a US National Institute of Standards and Technology (NIST) competition on hand-printed digits [67] 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-neighbour classification [58].

These observations suggest that large improvements in 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 programme may be expected to assist engineers by allowing

them to measure image quality, control the effects of variation in quality, and 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. Then, I summarize the recent history of image quality measurement 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 [8–10]. An earlier version was presented as an invited unpublished plenary address at the IAPR Workshop on Document Analysis Systems in Rio de Janeiro in 1999.

12.2 Document Image Degradations

By "degradations" (or, "defects") I 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.

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 such a page's contents can usually be considered 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 it should not be assumed 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 the present 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-colouring 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