ISSUES IN GROUND-TRUTHING GRAPHIC DOCUMENTS

Daniel Lopresti, Bell Labs, Lucent Technologies
George Nagy, Rensselaer Polytechnic Institute

> *And diff'ring judgements serve but to declare,*
> *That truth lies somewhere, if we knew but where.*
> - William Cowper 1731-1800

Is ground truth indeed fixed, unique and static? Or is it, like beauty, in the eyes of the beholder, relative and approximate? In OCR test datasets, from Highleyman's hand-digitized numerals on punched cards to the U-W, ISRI, NIST, and CEDAR CD-ROMs, the first point of view held sway. But in our recent experiments on printed tables, we have been forced to the second. This issue may arise more and more often as researchers attempt to evaluate recognition systems for increasingly complex graphic documents.

Strict validation with respect to reference data (i.e., ground truth) seems appropriate for pattern recognition systems designed for real applications where an appropriate set of samples is available. (The choice of sampling strategy for "real applications" is itself a recondite topic that we skirt here.) We examine the major components that seem to play a part in determining the nature of reference data. In the conventional scenario, the reference data is produced by entering some interpretation about each document using a chosen data-entry platform. Looking a little more closely at this process, we study its constituents and their interrelations:

Input format. The input represents the data provided to *both* interpreters and the recognition system. It is often a pixel array of optical scans of selected documents or parts thereof. It could also be in a specialized format: ASCII for text and tables collected from email, RTF or Latex for partially processed mainly-text documents, chain codes or medial axis transforms for line drawings.

Model. The model is a formal specification of the appearance and content of a document. A particular interpretation of the document is an instance of the model, as is the output of a recognition system. What do we do if the correct interpretation cannot be accommodated by the chosen model?

Reference Entry System. This could be as simple as a standard text editor like *vi* or *Notepad.* For graphic documents, some 2-D interaction is required. DAFS-1.0 (Illuminator), entity graphs, X-Y trees, and rectangular zone definition systems have been used for text layout. We used Daffy for tables. Questions that we will examine in greater detail are the *conformance* of the Reference Entry System to the Model (is it possible to enter reference data that does correspond to any possible model instance?), and its *bias* (does it favor some model instances over others?). To avoid discrepancies, should we expeditiously redefine the *Model* as the set of representations that can be produced by the reference entry system?

Verification and Reconciliation. Because the reference data should be more accurate than the recognition system being evaluated, the reference data is usually entered more than once, preferably by different operators, or even by trusted automated systems. Where there is a high degree of consensus, the results of multiple passes are *reconciled*. However, in more difficult tasks it may be desirable to retain several versions of the truth. We may also accept partial reference information. For instance, we may be satisfied with line-end coordinates of a drawing even if the reference entry system allows differentiating between leaders, dimension lines, and part boundaries. Isn't all reference data incomplete to a greater or lesser extent?

Truth format. The output of the previous stage must clearly have more information than the input. To facilitate comparison, ideally the ground-truth format is identical, or at least equivalent, to that of the output of the recognition system.

Personnel. For printed OCR, ordinary literacy is usually considered sufficient. For handwriting, literacy in the language of the document may be necessary. For more complicated tasks, some domain expertise is desirable. We will discuss the effects of subject matter expertise versus specialized training (as, for instance, for remote postal address entry, medical forms, or archival engineering drawing conversion). How much training should be focused on the model and the reference entry system versus the topical domain? Is consistency more important than accuracy? Can training itself introduce a bias that favors one recognition system over another?

Although each of these constituents plays a significant role in most reported graphic recognition experiments, they are seldom described explicitly. Perhaps there is something to be gained by spotlighting them in a situation where they don't play a subordinate role to new models, algorithms, data sets, or verification procedures.

As mentioned, we are interested in scenarios where the evaluation of an automated system is a quantitative measure based on automated comparison of the output files of the recognition system with a set of reference (“truth”) files produced by (human) interpreters from the same input. This is by no means the only possible scenario for evaluation. Several other methods, including the following, have merit.

1. The interpreter uses a different input than the recognition system (for example, hardcopy instead of digitized images).

2. The patterns to be recognized are produced from the truth files, as in the case of bitmaps of CAD files in GREC dashed-line or circular curve recognition contests. Do we lose something by accepting the source files as the unequivocal truth even if the output lends itself to plausible alternative interpretations?

3. The comparison is not automated: the output of the recognition system may be evaluated by inspection, or corrected by an operator in a timed session (a form of *goal directed evaluation*).

4. There is no reference data or ground-truth: in postal address reading, the number of undeliverable letters may be a good measure of the address readers’ performance.

The notion of ambiguity is not of course unique to pattern recognition. Every linguist, soothsayer and politician has a repertory of ambiguous statements. Perceptual psychologists delight in figure-ground illusions that change meaning in the blink of an eye. Motivation is notoriously ambiguous: “Is the Mona Lisa smiling?” We do not propose to investigate ambiguity for its own sake, but only insofar as it affects the practical aspects of evaluating symbol-based document analysis systems.

We provide examples from the literature and from our own experiments of non-trivial problems with each of the six major constituents of ground truth. Unless and until they are addressed, these problems appear to preclude the possibility of real progress in evaluating automated graphic recognition systems. For some of them, we can propose potential solutions.