Issues in Ground-Truthing Graphic Documents*

Daniel Lopresti
Bell Labs
Lucent Technologies, Inc.
600 Mountain Avenue
Murray Hill, NJ 07974 USA

George Nagy
Department of Electrical, Computer,
and Systems Engineering
Rensselaer Polytechnic Institute
Troy, NY 12180 USA

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

Abstract

Is ground-truth indeed fixed, unique and static? Or is it, like beauty, in the eyes of the beholder, relative and approximate? 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. We provide examples from the literature and from our own experiments where non-trivial problems with each of the components appear to preclude the possibility of real progress in evaluating automated graphic recognition systems, and propose possible solutions. Mostly, however, we raise far more questions than we currently have answers for.

Keywords: *document analysis, performance evaluation, ground-truth.*

1 Introduction

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 [11]. 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

*Presented at the *Fourth International Workshop on Graphics Recognition*, Kingston, Ontario, Canada, September 2001.

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 (see also Figure 1):

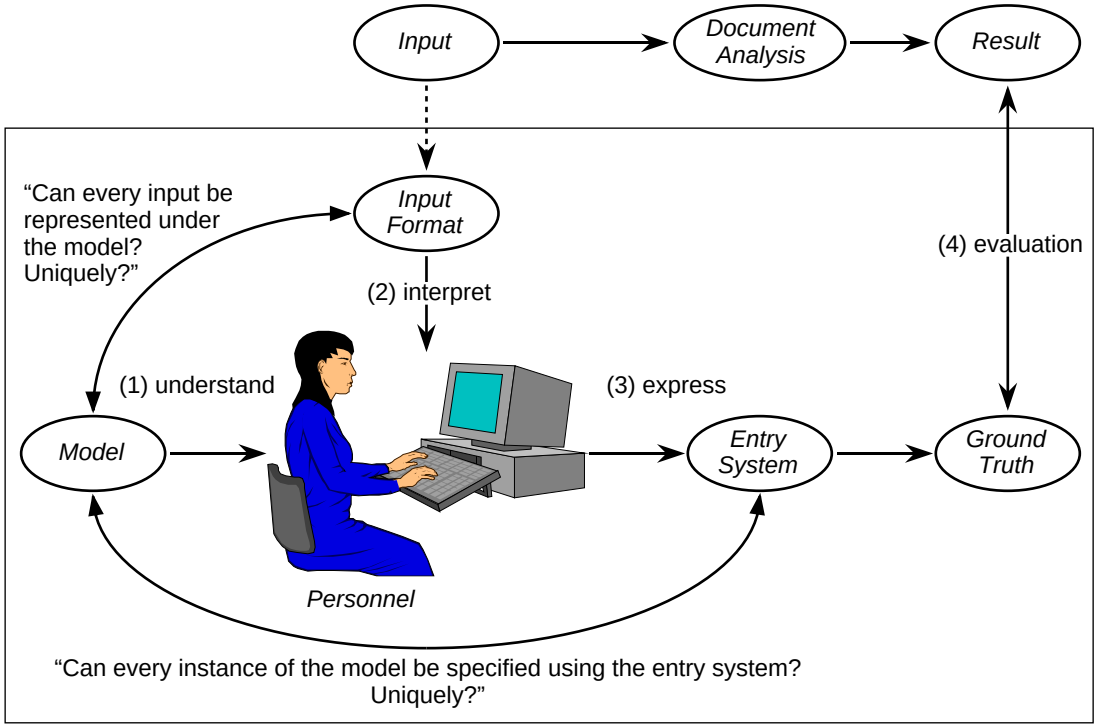

Figure 1: Overview of the ground-truthing process.

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 [5, 13, 18, 30]. We used our own system, Daffy, for tables. 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. 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?

Truth format. The output of the previous stage must clearly have more information than the input. To facilitate comparison, traditionally the ground-truth format is identical, or at least equivalent, to that of the output of the recognition system. 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. Is not all reference data incomplete to a greater or lesser extent? More flexible truth-formats may be needed.

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 do not play a subordinate role to new models, algorithms, data sets, or verification procedures.

2 The Role of Ground-Truth in Evaluation

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 [10, 12, 14], as in the case of bitmaps of CAD files in GREC dashed-line or circular curve recognition contests [3]. 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.

5. One approach used in medical image interpretation is double-blind evaluation [31]. In other words, the output of the automated system is simply included among outputs generated by human operators. The evaluators than rank the output (without reference to any specific ground-truth file).

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 (*e.g.*, Figure 2). 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.

Figure 2: A well-known figure/ground conundrum. How would you label this?

We provide examples from the literature and from our own experiments of non-trivial problems with each of the five 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.

3 Document Input Format

The natural input format for scanned documents is a bilevel, gray-scale, or RBG pixel array at some specified sampling resolution like 300 or 400 dpi. For electronic documents, there is ample choice. Documents intended for human consumption (printing or display) are often encoded in Postscript (PS) or Portable Document Format (PDF). At a given spatial sampling rate, these are lossless encodings and can be converted to bitmaps. PS and PDF are seldom used as input for document recognition experiments because the mapping to a pixel array is many-to-one.

How much format information is preserved by Hypertext Markup Language (HTML) depends on the specific encoding and on the browser. However, HTML can only represent layout and not semantic content, and is rapidly being replaced by XML. XML is only a meta-language: both the style of rendering and the semantic interpretation require ancillary information.

4 Model

What is a model? Who needs it?

The notion of models derives from the natural sciences, where they provide a concise explanation of observed facts and may predict yet unobserved phenomena. In DIA, we deal entirely with digital artifacts; therefore the models of interest are of a generative nature. Given an appropriate selection of input parameters, the model is expected to produce a digital object of the type being modeled. The model may be deterministic (blackboard systems, schema-based systems, syntactic methods, and procedurally coded rules [1]) or stochastic (Markov, Hidden Markov, stochastic context-free grammars).

Stochastic models can be used to generate synthetic data for experimentation. For instance, appropriately calibrated character defect models can produce multitudes of patterns, and relatively simple natural language models can produce multitudes of sentences. More complex models can generate valid addresses and short stories.

In experimentation with real data, models are used differently. An instance of a model that corresponds to a digital object may be viewed as a compressed representation that omits whatever is not essential for the purpose at hand. The object of the recognition system is to either recover the model parameters, or to declare that the given object is not a valid instance of the model.

The model is in some sense an idealized representation of the digital object. Therefore one could argue that imperfections of the object need not be represented in the model. So should we model, instead of the object, the intent of its maker? The perceptions of its readers?

For some automated data entry applications, a suitable model for a text region might be simply a string of character identities from a given alphabet. This was the model used in the ISRI OCR benchmarks, with the alphabet corresponding to the printable characters of the Latin-1 ASCII. The model could not represent bullets and Greek letters, nor two-dimensional symbol clusters like subscripts, superscripts, and mathematical formulas [23].

The Panasonic OCR experiments on *Moby Dick* followed similar conventions [4], which are more appropriate for literary, rather than technical, material.

Even this simple model raises questions. For examples, should typographic errors, like "hamnered" or "W00D" be corrected? This would reduce the error rate on natural-language text if the OCR system used language context, but would be quite misleading if the system were intended for correcting galley proofs.

For information retrieval, a better model might be a string of words from a pre-defined vocabulary. Most full-text search techniques can be implemented on such a model. NIST attempted to mount a large experiment to determine how much can "meta-data" (in this instance, additional format information) can improve query response [6, 7, 8], but the necessary experiments, combining OCR and IR, proved too difficult to execute.

Sequences of words are generally insufficient for postal addresses and forms. For such applications, it is desirable to model the types of acceptable fields (street address, delivery date), the permissible contents of fields (which are usually not enumerable), and the constraints between the field contents (like the presence of a city in a state). Such models can often be formalized as a specialized grammar. Note, however, that we would need a different model if (as is often the case) the 11-digit delivery-point zip-code was the recognition target [26].

For page reconstruction, symbol-based models are clearly inadequate. Knuth formulated the side-bearing model for typographic conventions based on pages composed from slugs and rulings inserted into a frame [15]. Alternatively, we may resort to a page-representation language like Rich Text Format (RTF), LaTeX, or proprietary word-processing code. However, the description of a page in any of these formats is not unique, and therefore it is difficult to use such models for page interpretation experiments. ScanSoft developed XDOC, which represents the location and size of the bounding box of each word, typeface, type size, and word transcription [25]. This model does not require character-level segmentation.

Such models allow checking the appearance of a reconstructed page more easily than a pixel-level model because they are insensitive to small changes in line or word location, and to subtle font substitutions. However, they correspond only to the lowest level of representation afforded by word processors, and neglect the logical labeling provided by style sheets. We may know that a one-line paragraph is in a large, bold font, but not that it is a section title. Modeling higher-level logical entities requires something like SGML or ODA, which separate form from function.

For tables, Wang developed an elegant model in the form of a formal data type [29]. The essence of this representation is that items in the cells of a table can be uniquely identified or addressed as the intersection of a set of categories (row and column headers). She found in her experiments that the greatest shortcoming of the model was its inability to deal with footnotes, which are often essential for correct table interpretation. Furthermore, the model does not take into account any ordering (*e.g.*, alphabetical or numerical) of items in a row or column. The model carries no semantics: for instance in a table of birth and death dates, lifetimes do not have to be positive. Wang's model does, however, have a rendering component that is completely separate from the logical representation.

Early experiments in equation recognition used LaTeX [19]. LaTeX provides a much weaker model than Maple or Mathematica because the Latex format does not allow evalu-

ating or simplifying expressions automatically.

Models for engineering drawings and circuit diagrams generally have provisions for representing the geometry and connectivity of the primitive entities. The model may also incorporate predefined symbols, and line-types and configurations such as dimension sets, center-lines, and hash-marks. Constraints may be introduced for ensuring the compatibility of orthogonal projections. It is, however, risky to restrict the model to valid objects or circuit, because there is no guarantee that the original drawing satisfies these constraints. Modeling cartographic maps raises similar problems. Only when there is good reason to have faith in the validity of the documents, as in the case of previously verified cadastral maps [2], should the model include all applicable constraints. It would obviously be futile to use PSPICE to verify the integrity of hand-drawn circuits if the model for digitization did not allow invalid circuits. If, on the other hand, the hand-drawn circuits are believed valid, then PSPICE could improve the conversion accuracy.

According to Blostein and Haken, modeling notational conventions is difficult because such notations are not formally defined but evolve by common usage. They claim categorically that "There are no ground-truth models of ideal generator output." (for diagram generation and recognition) [1].

5 Reference Entry System

The data entry system should ideally be able produce any valid instance of the specified model. If, for example, the model accommodates superscripts, then it is not enough to be able to enter the superscript above and to the right of the base symbol, but there must also be a way to indicate the asymmetric binary relation between the two symbols. If the model accommodates dot-dash lines, so must the reference entry system. If, in addition, the model is expressive enough to represent this line as the center line of the projection of a cylinder, than the entry system must also be able to record this relationship.

Taking advantage of the full power of the reference entry system can reduce the ambiguity of having many different representations of the same entity. For example, a dashed line with specified beginning and end points is easier to match then an unordered set of collinear dashes.

In practice, the time required for entering the ground-truth is of paramount importance. With current computer speeds and storage capacities, the effort required to prepare ground-truth is almost always the limiting factor on the size of experiments. The fastest systems for entering ground-truth are seldom those developed for DIA experiments. For most common applications – text, page layout, drawings, schematic diagrams, formulas, and music notation – efficient data entry and editing systems were developed long before attempts at automating the process. It is worthwhile, whenever possible, to make use of such legacy systems. (Blostein and Haken recommend that such systems should also be exploited for automated recognition [1].)

Manual data entry systems fall into two categories. The first category, like the early word processors, was designed only to produce pleasing printed output. These systems need not represent higher-level logical relationships, and therefore seldom do. The second category,

like spreadsheets, symbolic integrators, music editors, and circuit analysis packages, was intended from the first for computer analysis of the data. These systems, in addition to producing printable output, also offer means to represent logical relations, and are therefore more suitable for advanced DIA.

Reference entry systems often incorporate automated tools to reduce data entry errors. Examples are spelling and syntax checkers, difference detectors, and consistency verifiers.

6 Truth Format

As we have noted, for plain OCR, strings of alphabetic symbols, encoded in a standard form like ASCII or Unicode, is adequate. To obtain a unique string representation for each text zone, in the ISRI benchmarks of OCR, blank lines, and leading and trailing blanks on a line were eliminated, and consecutive blanks within a line were compressed to a single blank [23]. The University of Washington OCR database used TEXescape codes [21], while Bettels used SGML and a European Patent Office (EPO) proprietary character set to encode strings for multilingual patents [28]. Text zones are generally used because the string representation of the text portion of a whole page may not be unique unless the reading order is specified or inferred. The string representation is often modified to carry minimal 2-D information by means of tabs and carriage returns.

There is no widely accepted format for page layout, and every article that we have seen on the subject proposes its own taxonomy and file format for entities above the word level. At DARPA's request, RAF spent several years preparing a page representation language called Illuminator (originally "DAFS") for both OCR ground-truth and output [5], but it did not catch on widely.

Some researchers on tables have established their own format for representing the specific aspects of tables of interest in their research. Wang must have developed some file structures for her model. Abu Tarif proposed using Excel, while others suggested a DBMS. However, none of these capture the graphic qualities of tables.

For line drawings, one of the computer-automated drafting languages, like DXF, seems like an appropriate intermediate level format. The fact that DXL and other CAD languages may not be able to capture some aspects of hand-drawn drawings assumes less and less importance, as more engineering drawings are prepared using a computer.

7 Personnel

For preparing ground-truth for ordinary OCR, perhaps not even literacy, in the usually accepted sense of the term, is necessary. In the heydays of offshore document entry, we learned that non-English-speaking operators averaged higher speed and lower error rate than speakers of English! Page layout is another matter: many DIA researchers are not familiar with standard typographic jargon, and invent their own entities and terms. The preparation of ground-truth couched in these terms must therefore be performed by the research team.

For more complex documents some expertise is required. Even rows and columns are difficult to demarcate in foreign-language tables, and any interpretation is out of question. The complexity of table interpretation ranges from that of relatively common material, like stock-market quotes, baseball statistics, and railroad schedules, that are comprehensible to a large function of the public, to complex financial statements, tabulations of physical and chemical properties and experimental results, and reports of statistical tests, that can be understood only by specially trained personnel.

To create DXL-formatted ground-truth for engineering drawings, obviously a trained computer draftsman is required. However, an expert draftsman may capture the intent of the drawing, rather than what is actually scanned. By way of analogy to the foreign-language data entry operators, perhaps someone familiar with the graphical data entry system, but entirely naive with respect to the specified objects, should enter drawing ground-truth.

8 Intrinsic Ambiguity

Many of the questions we have raised concerning ground-truth have been visited before, directly or indirectly, by researchers who have preceded us. Every paper that reports experimental results must contend with the matter. In this section we survey briefly some of the larger-scale efforts where the development of ground-truth, and in particular issues of ambiguity and/or errors in the data, have played a primary role.

The well-known University of Washington UW1 CD-ROM database, for example, consists of over one thousand page images made available to the international community by Haralick and Phillips *et al.* [9, 20, 21, 22]. For the real scanned pages in the dataset (as opposed to the synthesized images), the ground-truth was carefully constructed using a double data entry, double reconciliation procedure. In other words, each page was interpreted independently by two different individuals. These data were then compared by two other individuals, each of whom attempted to correct any discrepancies between the two original interpretations. Then, in a final stage, a single person took the results from the two reconcilers and performed a final reconciliation. This care resulted in an error rate estimated to be extremely low (about 29 per million characters) and, as a result, UW1 has become one of the most-used resources in document analysis research. There was, however, a significant cost for this degree of rigor: each page required roughly six person-hours to produce.

It is interesting to note that the UW1 protocol is oriented towards preventing data entry errors and making sure there is a single, unique representation for every entity in the dataset. It does not allow for the possibility of multiple legitimate interpretations; the master reconciler is the final "arbiter." The zoning process – a task likely to lead to disagreement – was also handled by a single individual. Such policies are entirely appropriate for the kinds of applications UW1 was intended for, but may not be easily extensible to ground-truthing for more complex of document analysis tasks.

METTREC, the Metadata Text Retrieval Conference, was an effort planned by the National Institute of Standards and Technology (NIST) to evaluate document conversion

using OCR and its impact on information retrieval technologies [6, 7, 8]. This involved scanning over 67,000 pages from the *Federal Register* for 1994. Since the electronic text for each page was already available, the most labor-intensive parts of the ground-truthing process included scanning the page images, discarding blank pages, and making sure the images were correctly labeled as to which page they corresponded to (so they could be associated with the proper electronic text). This process was semi-automated by attempting to locate and read the page number field using OCR. The sort of ground-truth needed to evaluate text-based retrieval in this context is fairly basic – the page is treated as a "bag of words" with no real need to represent higher-level structure. Even so, and despite the fact that the basic steps seem relatively straightforward, the sheer quantity of pages made the task challenging.

In their paper describing the TRUEVIZ ground-truthing system, Kanungo *et al.* acknowledge that "a protocol for ground-truthing documents" is key, but this is not one of the issues they address in the work [13]. TRUEVIZ is a portable tool written in Java and uses XML to encode ground-truth markup. It offers multilingual support via Unicode and provides custom input methods to extend Java in this regard. It does not seem to address the potential for alternate interpretations, however.

Miyao and Haralick's paper on ground-truthing for music states that one of the requirements for ground-truth is "... whether or not the meaning of each piece of music can be interpreted correctly and whether or not the music can be correctly converted into playable data" [17]. Their current ground-truth database is small (eight sheets), and the paper does not consider the question of how to evaluate recognition results relative to the truth. Musical notation is, of course, rife with ambiguity – the artist's interpretation is central to a performance – so this would seem to be an application where the notion of a single, well-defined ground-truth may be hard to defend.

We have uncovered a number of papers that are beginning to recognize intrinsic ambiguity and its impact on the truthing process. In explaining problems in ground-truthing for the classification of audio segments, Srinivasan *et al.* observe [27]: "Even the ground truth definition can be subjective, for example, a clear speech segment followed by mixed audio which includes speech tends to be classified as a single speech segment. Analogously, a clear music segment followed by mixed audio which includes music tends to be classified as a single music segment." Generally speaking, indexing information required to group objects into categories is more sensitive than content information that affects "only" the specific interpretation of the object.

In a paper describing the Pink Panther system for ground-truthing and benchmarking page segmentation, Yanikoglu and Vincent state [30]: "Creating segmentation ground-truth is not always straightforward. A ground-truthing handbook was therefore written in which potentially ambiguous cases were described, in order to take as much subjectivity out of the ground-truthing process as possible. One of the decisions we have faced was the necessary level of detail; we have decided that, within limits, zones should be logically and structurally uniform, and that paragraphs were the smallest units of interest of regular text ..."

"Furthermore, in order to remove the main source of ground-truthing ambiguity, which is deciding on the label (subtype) of a region, we have expanded GroundsKeeper to allow multiple region labels for a given region. However, the benchmarking algorithm currently

uses only the first label of a region."

Even though they are addressing a problem from a fundamentally different domain, natural scene images versus scanned document pages, Martin *et al.* face exactly the same kinds of questions regarding interpretation, ambiguity, and perceptual refinement [16]. If the output from document analysis is intended to be what a human would perceive when he/she views the document (regardless of what is contained in the machine-readable encoding for the document), then perhaps there is not such a large difference between documents and natural scenes.

So far as we are aware, Martin and his colleagues may be among the first to examine the inherent variation in the truth created by different ground-truthers. While there seems on the surface to be significant variation, they argue that most of this is due to perceptual refinement and therefore not significant. They describe an evaluation measure that can factor out such differences, and conclude that there is indeed consistency between human ground-truthers, at least for the application in question (segmenting natural scenes) [16].

Recall, however, Figure 2 and its inherent ambiguity. While ground-truthers may very well agree on the locations of the edges in this image, could they ever agree on a labeling for the regions (without being forced in some way)?

Returning to the "easiest" problem we face in document analysis, the interpretation of character patterns, some interesting (and entertaining) examples of ambiguity can be found in Chapter 3 of the book by Rice *et al.* [24].

9 Issues and Open Questions

We admit readily that some of the notions that have been presented are an attempt to generalize to graphics the lessons that we have learned from text and table interpretation. In this section we attempt to collect together some of the questions we have raised throughout the earlier parts of this work in addition to some new ones. Our intention is that these might serve as the basis for future research investigations, or at least an interesting debate.

Is it sufficient to have just one version of the ground-truth when the input admits more than one interpretation? Provided the designated ground-truth is acceptable, this could never help a bad algorithm look better, but it could unduly penalize a good algorithm. If, however, there is debate as to whether the ground-truth is correct, that could be a problem. Who is to decide whether a ground-truth is correct? (Do we need a fair and impartial "jury" for ground-truth?)

What kinds of experiments could be used to confirm that a proffered ground-truth is appropriate for the task at hand? In other words, how can one tell whether uncertainties in the ground-truth will have a significant effect in the evaluation of an automated recognition system?

Context clearly plays an important role in human interpretation. What is the minimal context necessary for a ground-truthing task? What is the relationship between context-dependent interpretation and domain expertise? How does expectation fit in? For example, consider showing a ground-truther the following sequence of words:

```
TREE ... CAT ... DOG ... LETTUCE ... PEAR ... FOOD
```

He/she will certainly record that the final word is "food," even though it was written above using two zeros instead of "ohs."

On the other hand, if we show the same ground-truther the following sequence:

```
T65E ... C3T ... D7G ... L23472E ... P23R ... F00D
```

there is a much better chance he/she will identify the zeros as zeros.

Is the following "position," then, correct? The system is right when it arrives at an interpretation that some human might reasonably arrive at. It is wrong when it arrives at an interpretation that no human would arrive at. This is opposed to the current paradigm which states that the system is right when it arrives at the same interpretation that a specific individual arrived at when he/she created the ground-truth.

The issue of interpretation is surely central in ground-truth, since that is the final goal of DIA. Since DIA is generally less powerful than a human interpreter, it is likely to result in more ambiguity in interpretation. One operational question then is whether the ground-truther should also leave ambiguity, or whether the evaluation method should take into account of ambiguous automated interpretation vs clear human interpretation. At the very least, perhaps the ground-truth, like the output of many automated recognition systems should include some confidence measure that can be used in the evaluation criteria.

Do elements of the ground-truth have to be labeled with regard to their importance in some intrinsic hierarchy? Since the relative importance of errors depends on the ultimate use of the document analysis results, this suggests that ground-truth is application dependent. This will be increasingly the case as higher level information, containing less redundancy, is extracted from the document. For example, it may be less costly to mistake a a dashed line crossing in a drawing for a plus sign than to label a cross-section as "drive-chain detail" instead of "brake assembly." We have barely touched on the issue of the metric used to compare the automated recognition result with the ground-truth(s), and on the cost-accounting for errors.

It appears that with ground-truth we have a very similar question to recognition of unsegmented objects. We cannot do the segmentation (layout analysis) without interpreting the objects, and it may be hard to do the interpretation without segmentation. However, humans can interwine the two easily. We may therefore propose that we do not create detailed ground-truth at all, but only high-level database-type truth. (*e.g.* proposing to use only zip codes as ground-truth for a whole postal address [26]).

References

[1] D. Blostein and L. Haken. Using diagram generation software to improve diagram recognition: A case study of music notation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-21(11):1121–1136, November 1999.

[2] L. Boatto, V. Consorti, M. D. Bueno, S. D. Zenzo, V. Eramo, A. Esposito, F. Melcarne, M. Meucci, A. Morelli, M. Mosciatti, S. Scarci, and M. Tucci. An interpretation system for land register maps. *IEEE Computer*, 25(7):25–33, July 1992.

[3] A. K. Chhabra and I. Phillips. The Second International Graphics Recognition Contest – raster to vector conversion: A report. In K. Tombre and A. K. Chhabra, editors, *Graphics Recognition: Algorithms and Systems*, volume 1389 of *Lecture Notes in Computer Science*, pages 390–410. Springer-Verlag, Berlin, Germany, 1998.

[4] J. Esakov, D. P. Lopresti, J. S. Sandberg, and J. Zhou. Issues in automatic OCR error classification. In *Proceedings of the Third Annual Symposium on Document Analysis and Information Retrieval*, pages 401–412, Las Vegas, NV, April 1994.

[5] T. Fruchterman. DAFS: A standard for document and image understanding. In *Proceedings of the Symposium on Document Image Understanding Technology*, pages 94–100, Bowie, MD, October 1995.

[6] M. D. Garris. Document image recognition and retrieval: Where are we? In *Proceedings of Document Recognition and Retrieval VI (IS&T/SPIE Electronic Imaging)*, volume 3651, pages 141–150, San Jose, CA, January 1999.

[7] M. D. Garris, S. A. Janet, and W. W. Klein. Federal Register document image database. In *Proceedings of Document Recognition and Retrieval VI (IS&T/SPIE Electronic Imaging)*, volume 3651, pages 97–108, San Jose, CA, January 1999.

[8] M. D. Garris and W. W. Klein. Creating and validating a large image database for METTREC. Technical Report NISTIR 6090, National Institute of Standards and Technology, January 1998.

[9] J. Ha, R. M. Haralick, S. Chen, and I. T. Phillips. Estimating errors in document databases. In *Proceedings of the Third Annual Symposium on Document Analysis and Information Retrieval*, pages 435–459, Las Vegas, NV, April 1994.

[10] J. D. Hobby. Matching document images with ground truth. *International Journal on Document Analysis and Recognition*, 1(1):52–61, 1998.

[11] J. Hu, R. Kashi, D. Lopresti, G. Wilfong, and G. Nagy. Why table ground-truthing is hard. In *Proceedings of the Sixth International Conference on Document Analysis and Recognition*, Seattle, WA, September 2001. To appear.

[12] T. Kanungo and R. M. Haralick. An automatic closed-loop methodology for generating character groundtruth for scanned documents. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-21(2):179–183, February 1999.

[13] T. Kanungo, C. H. Lee, J. Czorapinski, and I. Bella. TRUEVIZ: a groundtruth / metadata editing and visualizing toolkit for OCR. In *Proceedings of Document Recognition and Retrieval VIII (IS&T/SPIE Electronic Imaging)*, volume 4307, pages 1–12, San Jose, CA, January 2001.

[14] D.-W. Kim and T. Kanungo. A point matching algorithm for automatic generation of groundtruth for document images. In *Proceedings of the Fourth IAPR International Workshop on Document Analysis Systems*, pages 475–485, Rio de Janeiro, Brazil, December 2000.

[15] D. Knuth. *The Texbook*. Addison-Wesley, 1984.

[16] D. Martin, C. Fowlkes, D. Tal, and J. Malik. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2001. To appear.

[17] H. Miyao and R. M. Haralick. Format of ground truth data used in the evaluation of the results of an optical music recognition system. In *Proceedings of the Fourth IAPR International Workshop on Document Analysis Systems*, pages 497–506, Rio de Janeiro, Brazil, December 2000.

[18] G. Nagy and S. Seth. Hierarchical representation of optically scanned documents. In *Proceedings of the Seventh International Conference on Pattern Recognition*, pages 347–349, Montréal, Canada, July 1984.

[19] M. Okamoto and A. Miyazawa. An experimental implementation of a document recognition system for papers containing mathematical expressions. In H. S. Baird, H. Bunke, and K. Yamamoto, editors, *Structured Document Image Analysis*. Springer-Verlag, Berlin, Germany, 1992.

[20] I. Phillips, J. Ha, R. Haralick, and D. Dori. The implementation methodology for the CD-ROM English document database. In *Proceedings of Second International Conference on Document Analysis and Recognition*, pages 484–487, Tsukuba Science City, Japan, October 1993.

[21] I. T. Phillips, S. Chen, J. Ha, and R. M. Haralick. English document database design and implementation methodology. In *Proceedings of the Second Annual Symposium on Document Analysis and Information Retrieval*, pages 65–104, Las Vegas, NV, April 1993.

[22] I. T. Phillips, J. Ha, S. Chen, and R. M. Haralick. Implementation methodology and error analysis for the CD-ROM English document database. In *Proceedings of the AIPR Workshop*, Washington DC, October 1993.

[23] S. V. Rice, J. Kanai, and T. A. Nartker. Preparing OCR test data. Technical Report TR-93-08, UNLV Information Science Research Institute, Las Vegas, NV, June 1993.

[24] S. V. Rice, G. Nagy, and T. A. Nartker. *Optical Character Recognition: An Illustrated Guide to the Frontier*. Kluwer Academic Publishers, Norwell, MA, 1999.

[25] ScanSoft, Inc., Peabody, MA. *XDOC Data Format, Technical Specification Version 3.0*, May 1997.

[26] S. Setlur, V. Govindaraju, and S. Srihari. Truthing, testing and evaluation issues in complex systems. In *Proceedings of the Symposium on Document Image Understanding Technology*, pages 131–140, Annapolis, MD, April 2001.

[27] S. Srinivasan, D. Petkovic, and D. Ponceleon. Towards robust features for classifying audio in the CueVideo system. In *Proceedings of ACM Multimedia '99*, pages 393–400, Orlando FL, November 1999.

[28] H. R. Stabler. Experiences with high-volume, high accuracy document capture. In A. L. Spitz and A. Dengel, editors, *Document Analysis Systems*, pages 38–51. World Scientific, Singapore, 1995.

[29] X. Wang. *Tabular abstraction, editing, and formatting.* PhD thesis, University of Waterloo, 1996.

[30] B. A. Yanikoglu and L. Vincent. Pink Panther: a complete environment for ground-truthing and benchmarking document page segmentation. *Pattern Recognition*, 31(9):1191–1204, 1998.

[31] T. S. Yoo, M. J. Ackerman, and M. Vannier. Towards a common validation methodology for segmentation and registration algorithms. In S. Delp, A. DiGioia, and B. Jaramaz, editors, *Medical Image Computing and Computer-Assisted Intervention*, volume 1935 of *Lecture Notes in Computer Science*, pages 422–431. Springer-Verlag, Berlin, Germany, 2000.