Scientific Challenges Underlying Production Document Processing

Eric Saund
Perceptual Document Analysis
Intelligent Systems Laboratory
Palo Alto Research Center
3333 Coyote Hill Rd.
Palo Alto, CA 94304
{saund@parc.com}

ABSTRACT

The Field of Document Recognition is bipolar. On one end lies the excellent work of academic institutions engaging in original research on scientifically interesting topics. On the other end lies the document recognition industry which services needs for high-volume data capture for transaction and back-office applications. These realms seldom meet, yet the need is great to address technical hurdles for practical problems using modern approaches from the Document Recognition, Computer Vision, and Machine Learning disciplines. We reflect on three categories of problems we have encountered which are both scientifically challenging and of high practical value. These are Doctype Classification, Functional Role Labeling, and Document Sets. Doctype Classification asks, "What is the type of page I am looking at?" Functional Role Labeling asks, "What is the status of text and graphical elements in a model of document structure?" Document Sets asks, "How are pages and their contents related to one another?" Each of these has ad hoc engineering approaches that provide 40-80% solutions, and each of them begs for a deeply grounded formulation both to provide understanding and to attain the remaining 20-60% of practical value. The practical need is not purely technical but also depends on the user experience in application setup and configuration, and in collection and groundtruthing of sample documents. The challenge therefore extends beyond the science behind document image recognition and into user interface and user experience design.

Keywords: doctype classification, functional role labeling, document sets, document image understanding

1. INTRODUCTION

Machine reading of document images was one of the first envisioned practical everyday applications for electronic computers and indeed by the 1980s Optical Character Recognition became the first large scale consumer and commercial application for pattern classification. By some standards, Document Recognition can be considered a relatively mature field. OCR is bundled with every scanner, libraries of printed newspapers and books are now available online, and literally billions of characters are recognized from scanned documents every day.

Yet, many daunting challenges confront this field. Recognition of handwriting and degraded images is still problematical. Non-roman scripts pose difficulties beyond the character segmentation and classification approaches developed for Western languages. And layout analysis remains an unsolved problem.

Thirty-five years after its conception, the vision of the paperless office has become reality as many classes of documents are now created, stored, transmitted, and read purely in electronic form. Yet document recognition is far more than character classification, and the need for this technology will persist into the foreseeable future. One definition for "document" is, "information presented in a format for human reading"; therefore, the image remains the common denominator for communication even in the midst of the semantic web and electronic data transfer. Much business is still conducted via paper. And if not paper, then pdf, Word, and other document formats fail to specify metadata cues adequate to interpreting meaning among the renderable glyphs they carry.

In this context, we observe some degree of dichotomy between academic research in document recognition, and what is one of the largest applications of document recognition technology—production document processing.

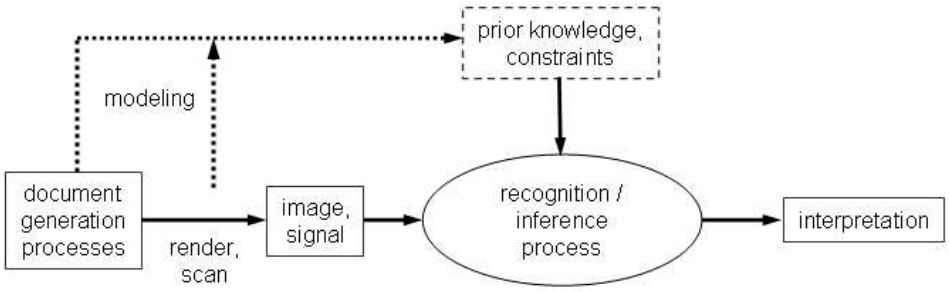

Figure 1. The classical paradigm for image and signal interpretation applies to document recognition. A signal arises from some causal processes in the world. Specifically in our case these are document image generation processes in all aspects and forms including authoring, composition, drafting, printing, typing, hand-annotating, paper folding, coffee spilling, scanning, faxing, etc. By modeling these, we obtain prior knowledge which makes inference possible by constraining interpretations of the document image.

Perusal of recent conferences and journal publications reveals a strong focus on reading of multi-lingual text, on historical documents, on segmentation of different types of markings, on handwriting, and on layout analysis of documents normally encountered by high-skilled knowledge workers, such as articles and business correspondence. On the other hand, a large industry has grown around data capture from transaction documents such as forms, purchase orders, invoices, loans, medical claims, and customer correspondence. Millions of transaction documents are created, and therefore need to be read, every day. The demand for automation is therefore very large. Transaction documents are not alien to academic study, but they are seldom a focus. Industry representation at academic conferences is mainly limited to a few OCR companies. At industrial trade shows one seldom sees anything like a scientifically motivated presentation or forthright appreciation of the technical challenges underlying the products and services that are enthusiastically promoted.

This paper attempts to highlight scientifically interesting research problems that fall in this gap. We focus on problems encountered in high-volume production document processing. Typically the business task is called "data entry." Traditionally this has been done by armies of workers sitting at consoles and typing in information presented to them on a screen. The internet has enabled outsourcing of data entry tasks worldwide in pursuit of low labor costs. Nonetheless, even greater increases in productivity are promised by automation—if only computers could become as capable at reading documents as the human visual system is. It is easy from a distance to regard data entry as a routine and therefore readily programmable task. But the visual intelligence required to read documents is as sophisticated as that to recognize faces, monitor vehicle traffic, direct robots, or perform other potential jobs that lie at the edges of the state of the art of computer vision. The need and opportunities for research are manifest.

We organize this evaluation according to a problem decomposition that has been established in the commercial world, namely *doctype classification*, *data capture*, and a third class of problem that has not yet emerged commercially but we believe is coming soon, which we call *document sets*. Doctype classification is assigning a page image to a predefined category. Data capture is about reading pertinent information expected to be found on the document page; the general problem can be called *Functional Role Labeling*. Document Sets is about relating pages and their contents to one another according to business logic. All of these topics stand to benefit from new developments and scientific methodologies brought by academic traditions.

In all cases, we observe that the classical paradigm of computer vision and signal analysis applies,[1] as depicted in Figure 1: An unknown signal is interpreted in terms of models for signal structure and task output. Inference is enabled by prior knowledge of the signal source. If the document to be recognized is strongly constrained, then strong prior models can be applied and the inference of task-specific output can be readily formulated. But source variability and noise in the input brings uncertainty to interpretations and complexity to the use of prior knowledge, which in turn demands more flexible and sophisticated inference methods. This conundrum is most clearly seen first in the task of doctype classification.

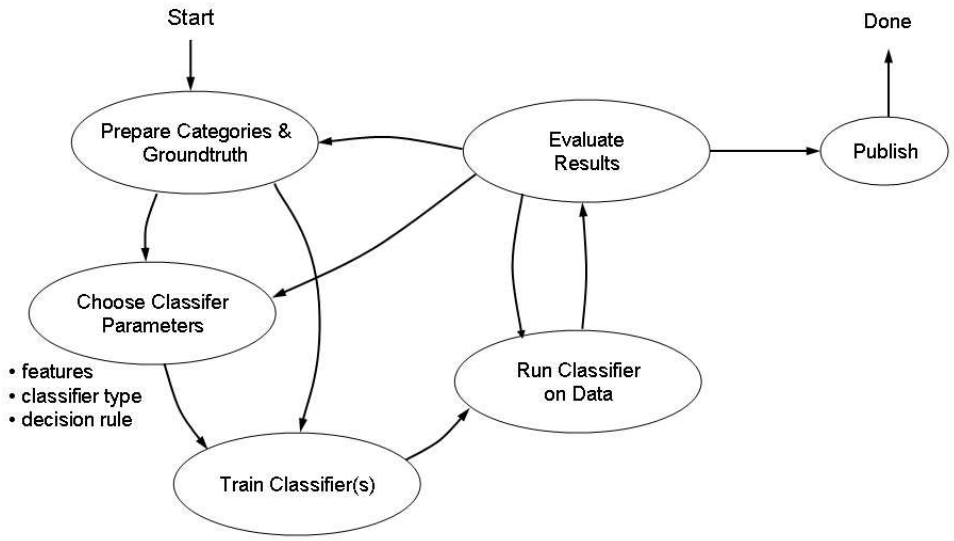

Figure 2. Workflow process for configuring a doctype classification job.

2. DOCTYPE CLASSIFICATION

Doctype classification in its simplest form is the assignment of a document to one of a set of known categories, or types. This is simpler if the categories are very narrowly defined such as known forms like "2010 Tax Form 1040 Side 1", "AMA Dental Form J515", "Change of beneficiary form 101954HL (05/09)". It is more difficult if the categories are loosely defined, e.g. "Drivers' License", "Invoice", or "Correspondence requesting cancellation of service." Current and future academic literature may address the deeper conceptual issues surrounding the rather open-ended considerations of document genre analysis—like people, documents come with all sorts of properties and organizational structures defying simple categorical binning. But for purposes of production document processing we may assume that doctype classification is performed for some well-defined business function for which some crisp specification of doctype categories has been established.

The dominant method for practical doctype classification in the commercial world is image template matching. When the doctype is a known form or otherwise highly constrained, then certain indicative image regions such as a logo or consistent text label can be fairly reliable. Sometimes multiple templates are applied with combination of threshold match scores and AND and OR rules.

A second common doctype classification approach employs feature extraction and classifiers trained by machine learning. Features may include Fourier components, HAAR-like filter responses, and word counts (following OCR). Sometimes dimensionality reduction such as Principal Components is applied. Classifiers include near-neighbor, decision trees, Support Vector Machines, and generative density estimators.[2]

Each of these approaches has strengths and weaknesses for practical production document processing, and there remain many opportunities for additional research. One set of difficulties for rule-based approaches is writing and verification of the rules. For example if a given doctype contains a fixed logo 95% of the time, then an image template matching the logo could work roughly 95% of the time but some other template or templates must cover the remaining 5% of cases. This leads to a cumbersome process of choosing templates, authoring decision logic, then testing and verifying the process.

Another set of complexities arises with machine-learning based approaches. The feature set must be selected or designed, the classifier method must be chosen, and its parameters tuned. Then a large enough training set must be groundtruthed, a subset selected and used for training, and the remaining subset used for cross-validation.

Neither approach is entirely satisfactory for the circumstances of commercial deployment of doctype classification in production document processing systems. Both entail a great deal of manual effort. Semi-skilled labor

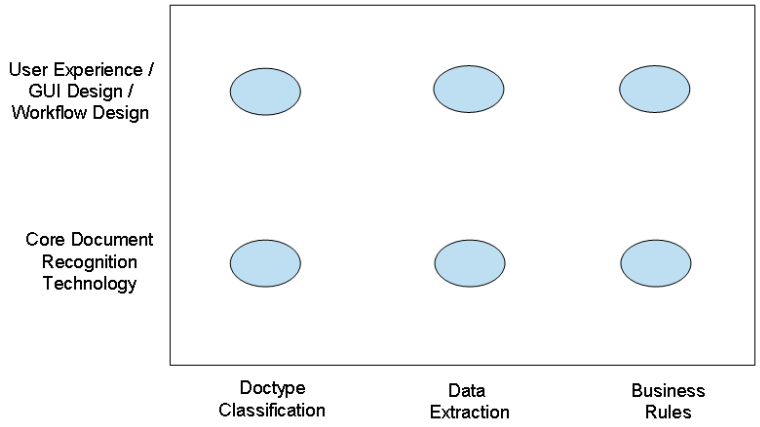

Figure 3. Three major categories of production document processing phases require attention at both the core technology and user interface/user experience design levels.

is required to obtain sufficient groundtruthed exemplars required to validate rules and train and test machine learning approaches. Most machine learning algorithms require hundreds, or at best, a few tens of exemplars per category. As a practical matter, these are not so easy to obtain, often because the job must be partially configured and bid according to sparse customer specifications before a full-fledged production configuration is developed. Then, the selection of templates and rules, or feature and classifier parameters, requires concentrated attention of skilled technicians and engineers. Common experience for off-the-shelf commercial products is that the per-document doctype classification cost is preceded by expensive consulting services on the part of the vendor to set up each job.

These observations point to two issues for research attention, namely few-exemplar learning, and user experience design for doctype classifier configuration.

Typically when a production doctype classification job is described to the setup technician, the target doctypes are provided in terms of single exemplar documents. Clients don't appreciate that computer algorithms need explicit examples of allowable variations in what to a person are clearly the same doctype. Distinctions among categories are obvious to human observers because we implicitly bring background knowledge and experience informing the distinctive qualities that define the categories. Current rule-based or machine learning-based paradigms lack this knowledge so must be programmed or trained in exquisite detail. If a large body of prior knowledge about document appearance, spectrum of textual contents, and layouts could be constructed to empower a doctype classification system at the outset, this could perhaps simplify the process. At best, we would like to achieve one-shot learning, whereby a single exemplar would suffice to define each doctype category with all of its implicit characteristics distinguishing it from other categories for the application job.

A second and related issue pertains to the user interface provided to the setup technician for designing rules and templates, or for selecting features and operating the training, cross-validation, and testing phases of machine learning paradigms. These steps follow a workflow process reflected in Figure 2. The design challenge is to provide a user interface that guides the technician through these steps, and helps them select useful choices where necessary.

User interface design interacts with the core recognition technology underlying doctype classification. In general, more powerful recognition technology enables fewer, simpler steps in the configuration process. For example, if reliable one-shot learning were available at the core, then the groundtruthing would require few samples and cross-validation could become less rigorous. As depicted in Figure 3, both core technology and user interface aspects of the system-level solution are present across three categories of production document processing phases, doctype classification, data extraction, and application of business rules.

One example for enhancing core technology in order to simplify the configuration process is reflected in recent research on automated learning of image anchor templates for doctype classification. Sarkar has developed a

process for automatically selecting image templates that are maximally discriminative for classifying structured appearance doctypes, i.e. forms, amongst one another.[3] The process involves a generate-and-test approach over candidate image templates selected according to heuristics reflecting generic knowledge about typical properties of forms documents. Given this core technology, the task of the configuration technician is simplified to that of providing a handful of exemplars and initiating the automatic learning algorithm. The tedious process of selecting and testing templates by trial-and-error is eliminated.

3. DATA CAPTURE / FUNCTIONAL ROLE LABELING

The structure of production document processing systems is typically that, following doctype classification, the principal task is to read certain information from each document according to its doctype. If the doctype is a form, then specified data fields are read. If the doctype is an invoice, then in addition to expected fields such as vendor name, purchase order number, and total due, the task is to capture line item details about each of the items purchased such as item description, quantity, catalog number, per-item cost, and total cost. In general we call this process *Functional Role Labeling*, that is, the various textual items carry different information, or serve different functional roles, with respect to the meaning of the document. Most often, layout is the key to defining functional roles. *

Showcase examples of functional role labeling for non-transactional documents are found in work on attaching metadata to scanned or online journal articles. There, roles include title, author, keywords, abstract, etc. Recent work by Gao and Wang[5] illustrates a promising paradigm borrowed from speech recognition and handwriting recognition. This paradigm follows four stages:

(1) oversegment the signal to create atomic units

(2) group atoms to form hypotheses for meaningful larger structures

(3) score individual hypotheses according to desired criteria

(4) perform search to select combinations of hypotheses optimizing global criteria.

Document knowledge resides in the formulations of each of these stages for the purposes of document recognition tasks.

A second, very different, approach follows the case-based recognition paradigm. This involves matching input data to closest exemplars in a reference set. An excellent example is recent work in functional role labeling of medical journals by Chen et al.[6] Both of these approaches could in theory be adopted to transaction documents and thereby become applicable to production document processing.

As with doctype classification, the difficulty of data capture from forms depends on the degree of prior constraint that can be applied. In the case of known forms, the simplest approach is to rely on locations of data items according to stored templates. Known as "Zonal OCR," this requires proper alignment and scaling of the document image. Straightforward deskew and global registration methods sometimes achieve this. In practice, however, other complications often arise. It is common for data items to be typed, handwritten, or printed outside the bounds of their target zones. Data entries frequently overlap rule lines of forms, corrupting OCR. Several methods have been proposed for detecting rule lines and reconstructing broken characters. A challenge remains in disambiguating the jumble of markings that can occur in an imperfectly filled-out form. Causal sources of markings that should be distinguished include form graphic, form text, entered data, handwritten graphic, handwritten textual annotation, stamp, and image noise. See Figure 4. Recent work in pixel-level labeling would seem to be applicable[7–9] because fragmentation of foreground markings into units smaller than connected components is essential to this task, while much academic work has stopped at the connected component level.[10,11] The classic segmentation/recognition dilemma of mainstream computer vision is operative here: given correctly segmented image layers, OCR and graphics recognition would be much more tractable, and

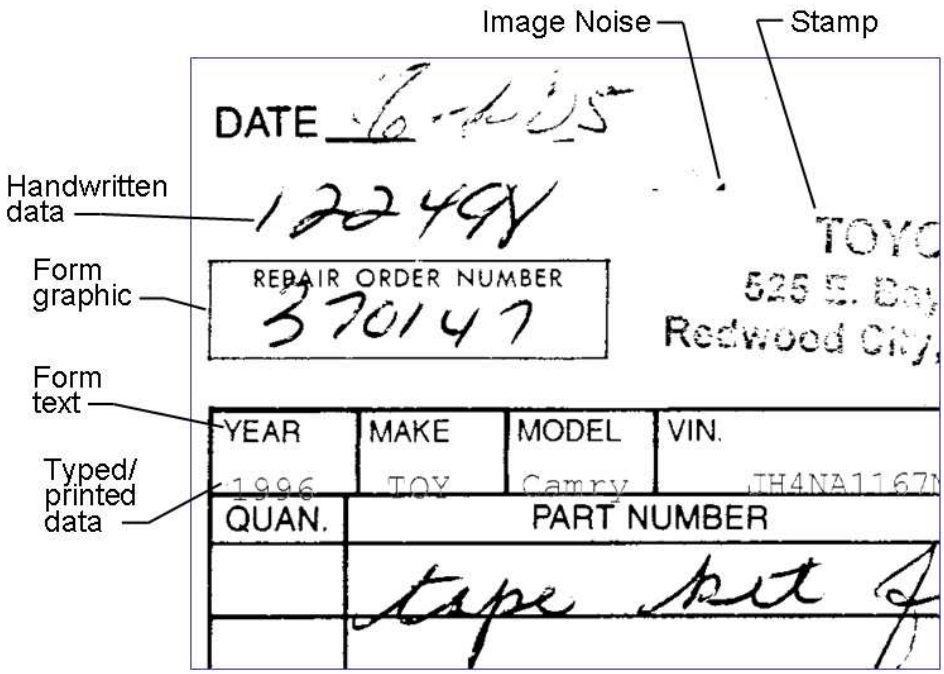

Figure 4. Production scale processing is made difficult by the mixture of markings found on forms documents.

conversely, if the text and graphical shapes were known in advance, it would be relatively easier to discover their poses and supporting pixel-level evidence in the image.

Functional role labeling becomes more difficult as document layout structure becomes less predictable. Invoices, for example, display myriad layout styles. Common industry practice is centered on rule-based approaches whereby data fields are identified through a combination of text pattern filtering and proximity to indicator fields. For example, an invoice payment due date would be declared if a text pattern resembling "Due Date", or "Payment due:" or "Pay by," etc., is found, and then, near below or to the right is found a text string fitting a "mm-dd-yyyy" pattern. Indicator fields and data value fields can take many forms. Often regular expressions are used to specify and match anticipated textual patterns. But these can be extremely tedious to write and test. Moreover, text pattern matchers must in many cases be made tolerant to OCR errors. Scientific advances in string pattern matching beyond manual regular expression authoring, and automatic string pattern induction from examples, would be most applicable here.[12]

In addition to text pattern content, the problem of functional role labeling has a spatial aspect that is subject to the layout style of the document. Figure 5 shows how proximity, direction, and graphical elements can vary in associating a *indicator* for the existence of an invoice number with a *data value* for the invoice number, in various invoices. A difficulty in creating rules for assigning functional roles to document entities is that rules applicable to one layout style will often fail for another. Style modeling in document recognition has been studied in the context of glyph appearance for OCR.[13] What is needed are new modeling frameworks that will support the learning and adaptation of document layout models to a diversity of styles without a great deal of labor-intensive and error-prone manual coding of rules, exceptions, tolerances, and thresholds.

A final example of functional role labeling in document images occurs with tabular data. Invoices, receipts, purchase orders, healthcare claims, and financial statements are transactional document genres that employ tabular organization of variable data. Rule-based approaches using line-art and whitespace separators work fine for simple tables whose data cells fall in neat grid. In practice, however, nominally tabular layouts frequently

*We formerly used the term, "Semantic Role Labeling"[4] but believe "Functional Role Labeling" to more precisely indicate that the central issue is functional roles in document composition and layout models.

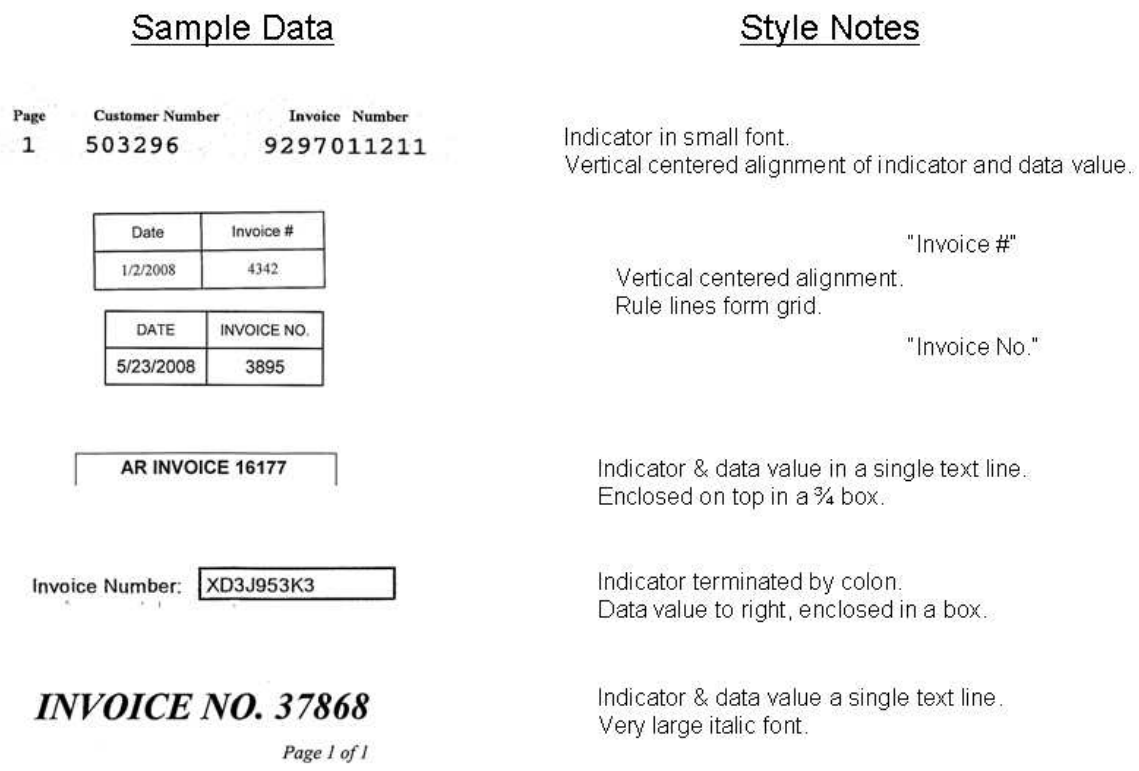

Figure 5. Prior knowledge required to effectively perform functional role labeling of invoice documents involves understanding the diversity of layout styles used in these documents.

display multi-line cells, cells that extend outside their formal boundaries, notations, explanations, and exceptional text, and tables that span multiple pages with intervening header and footer material. All of these phenomena drive us to seek methods possessing the kind of perceptual common sense a human reader brings to the document but are exceedingly difficult to specify through procedural programming methods.

One promising approach to this problem developed by Bart and Sarkar[14] is based on forming and scoring many hypotheses for layout structure, then performing efficient search for high-scoring hypotheses, as mentioned above. Hypotheses consist of assignments of words (obtained by OCR), to table cell definitions provided by a single labeled example. Scoring is a learned weighting of multiple heuristic factors such as visual alignment, proximity, and symmetry. Finally, search is conducted efficiently using a novel algorithm that provides efficient bounds-based pruning, called Best-First Leaf Search.

As with Doctype Classification, the task of setting up and configuring the Functional Role Labeling aspects of a data capture task can be labor intensive. The issues of user interface design and overall user experience in performing the full task are often overlooked in the academic document recognition community. Yet, just as Human Computer Interaction has become a first-class topic of study in Computer Science, the human factors involved in training and configuring underlying document recognition technology are in fact critical in achieving success in commercial applications. For example, if each and every invoice layout style and tabular data format must be hand-coded with tens of groundtruth training examples, the cost will impose an unacceptable drag on practical deployment. Ideally, the user would communicate their goals in interpreting a semi-structured document in an expressive modality similar to instructing a human reader/operator in going about capturing target data. What human-to-machine interaction lacks—given the state of current document recognition technology—that human-to-human interaction possesses, is common grounding in innate visual perception and culturally acquired knowledge about how to visually apprehend a document image. As machine document recognition technology improves, the remaining gulf can only be bridged by thoughtful and innovative UI design. Witnessing the energy

devoted to designing today's remarkable web sites and computer games, we believe that comparable attention brought to document recognition systems should achieve qualitatively remarkable advances in their practical utility.

4. DOCUMENT SETS

A third broad category of opportunity for research in document recognition is poised to emerge. This we may call the *document sets* problem. Document sets first pertain to documents spanning multiple pages, where information across the multiple pages can profitably be tied together. Beyond that, document sets pertains to collections of multiple single- and multi-page documents.

To date, the vast bulk of document recognition research has been directed to single pages. A notable exception is OCR performance enhancement through leveraging the repeated occurrences of glyphs across many pages.[15] A second example of multi-page document recognition is the cross-indexing of book title pages with chapters, as seen in Google's book indexing. These examples are only scratching the surface.

A much larger challenge is found in complex transaction documents such as tax returns, mortgage applications, healthcare claims, and financial accounting. In these document sets, pages and their constituent data items form complex webs of relationships. Names, dates, addresses, dollar amounts and their sums and percentages, signatures, and many other data entities are found throughout a document case. Sometimes, the practical task is to ensure that like items simply match one another. In other cases, the relationships among items are dictated by *business rules*. Business rules typically are logical expressions specifying nominal and exceptional conditions among entities in the collection. Here are some examples representative of real-life document processing jobs:

- If the deductions exceeds the standard deduction for this tax filer's filing status then the deductions must be itemized.
- If the account holder is not a U.S. citizen then a copy of their visa must be included.
- The application requires proof of income; this can be either of two forms: (1) pay statements for three successive months including the most recent month, or else (2) the applicant's W-2 form for the latest tax year but (2) applies only if the applicant has been employed at their present job for at least five years.
- All pages of the declarations document must be initialed by every applicant.
- Every document requiring a signature must be signed by every applicant. However, these signatures need not be found on the same pages; duplicate copies are acceptable as long as each party has signed at least one copy of each signature page.

Clearly, the consideration of business rules of this nature can be very labor-intensive, but it falls far outside the scope of published academic research in document recognition. Perhaps the closest consideration lies in the problem of citation rectification.[16] The citation rectification problem addresses the fact that applicable logical rules can be ill-specified and subject to uncertainty at the signal measurement level. The way in which a name, publication venue, institution, document title, or date is expressed is subject to great stylistic variation and ambiguity in spelling, abbreviation, and omission. Similarly, in transaction documents, many data entities can be ambiguous or problematical to collect at the document reading level. After all, the application of business rules relies on doctype classification and functional role labeling, and as we have seen, these are problematical steps in their own right.

Therefore, we believe that any appropriate formulation for the interpretation of business rules in the analysis of document sets must allow for uncertainty, ambiguity, and multiplicity of possible interpretations in view of prior models of documents and the task domain under which they are interpreted. The challenge to the document recognition field is formidable, for indeed it calls upon the full bearing of the encompassing endeavor of Artificial Intelligence.

5. CONCLUSION

Societal progress in our technology-driven world depends critically on advances in science and their adoption in real-world applications.[17] The practical problems of production document processing beg for attention by the corresponding fields of academic study.

Perhaps the biggest barrier to bridging these realms lies in communicating the nature of the problems in detail. This depends critically on a supply of sample data on which to base theory, experiments, and prototype solutions. Here lies perhaps the knottiest problem of all. Despite the millions of transactional documents processed in industry every day, formidable barriers stand in the way of sharing these for purposes of academic study. Often, documents contain confidential personal or proprietary data that cannot be shared by law or by contractual considerations. For example, automated capture of data from medical records and claims would be of huge societal value, but release of such data quite properly falls under severe legal restrictions under HIPPA statues. This issue was addressed at one point in the 1990's by the United States' National Institutes of Standards and Technologies benchmark Special Database Collections simulating U.S. Tax forms.[18] Comparable data sets reflecting doctype classification, functional role labeling, and document sets problems in transaction document processing are desperately needed if academically motivated intellectual resources are effectively to be brought to bear.

REFERENCES

[1] Witkin, A. and Tenenbaum, M., "On the role of structure in vision," in [*Human and Machine Vision*], Beck, J., Hope, B., and Rosenfeld, A., eds., 481–543, Academic Press, Orlando (1983).

[2] Sarkar, P., "Image classifcation: Classifying distributions of visual features," *18th International Conference on Pattern Recognition (ICPR 2006)* (2006).

[3] Sarkar, P., "Learning image anchor templates for document classification and data extraction," *20th International Conference on Pattern Recognition (ICPR 2010)* (2010).

[4] Sarkar, P. and Saund, E., "Perceptual organization in semantic role labeling," *2005 Symposium on Document Image Understanding Technology (SDIUT 2005)* (2005).

[5] Gao, D., Wang, Y., Hindi, H., and Do, M., "Decompose document image using integer linear programming," *International Conference on Document Analysis and Recognition (ICDAR 2007)* (2007).

[6] Chen, S., Mao, S., and Thoma, G., "Simultaneous layout style and logical entity recognition in a heterogeneous collection of documents," *Proc. International Conference on Document Analysis and Recognition (ICDAR 2007)* , 118–122 (2007).

[7] Moll, M., Baird, H., and An, C., "Truthing for pixel-accurate segmentation," *Eighth IAPR International Workshop on Document Analysis Systems (DAS 2008)* , 379–385 (2008).

[8] Sarkar, P., Saund, E., and Lin, J., "Classifying foreground pixels in document images," *International Conference on Document Analysis and Recognition (ICDAR 2009)* (2009).

[9] Saund, E., Lin, J., and Sarkar, P., "Pixlabeler: User interface for pixel-level labeling of elements in document images," *International Conference on Document Analysis and Recognition (ICDAR 2009)* (2009).

[10] Zheng, Y., Li, H., and Doermann, D., "Machine printed text and handwriting identification in noisy document images," *IEEE Trans. Pattern Analysis and Machine Intelligence* **Vol. 26, No. 3** (2004).

[11] Peng, X., Setlur, S., Govindaraju, V., Sitaram, R., and Bhuvanagiri, K., "Markov random field based text identification from annotated machine printed documents," *10th International Conference on Document Analysis and Recognition (IJDAR 2009)* (2009).

[12] Gulan, S. and Fernau, H., "An optimal construction of finite automata from regular expressions," *Proc. Foundations of Software Technology and Theoretical Computer Science (FSTTCS 2008)* , 211–222 (2008).

[13] Sarkar, P. and Nagy, G., "Style consistent classification of isogenous patterns," *IEEE Transactions on Pattern Analysis and Machine Intelligence* **27**, 88–98 (January 2005).

[14] Bart, E. and Sarkar, P., "Information extraction by finding repeated structure," *Ninth IAPR International Workshop on Document Analysis Systems* (2010).

[15] Xiu, P. and Baird, H., "Towardswhole-book recognition," *Proc. IAPR Int'l Workshop on Document Analysis Systems (DAS 2008)* (2008).

[16] Culotta, A., Wick, M., Hall, R., Marzilli, M., and McCallum, A., "Canonicalization of database records using adaptive similarity measures," *Thirteenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2007)* (2007).

[17] Arthur, B., [*The Nature of Technology: What It Is and How It Evolves*], Free Press, New York (2009).

[18] Dimmick, D., Garris, M., and Wilson, C., "Structured forms database," Tech. Rep. Technical Report Special Database 2, SFRS, National Institutte of Standards and Technology (December 2001).