The Future of Document Imaging in the Era of Electronic Documents

Thomas M. Breuel

Dept. of Computer Science
U. of Kaiserslautern
Kaiserslautern
Germany

DFKI
Erwin Schroedinger Str.
67608 Kaiserslautern
Germany

Abstract

Document imaging and document analysis are technologies for the interpretation and manipulation of document images. It is commonly assumed that the increased use of electronic documents and data communications will obviate the need for document imaging and document analysis, as more and more documents are exchanged in formats like HTML, XML, PDF, and other well-defined, structured formats. This paper examines the question of how likely it is that paper will be replaced by electronic documents in the near future, what possibilities exist for paper and electronic documents to co-exist, and what role document imaging and document analysis will play as electronic communications and computers become ever more widespread and portable.

Introduction

Large numbers of documents are now authored with computers using software like email systems, web authoring tools, word processors (Microsoft Word, OpenOffice), and markup languages (LaTeX, SGML, DocBook). Furthermore, documents are commonly exchanged electronically, over the world-wide web (news articles, scientific publications, government publications, business forms, product brochures, etc.). They are distributed and exchanged via email systems, as well as via groupware and document repositories.

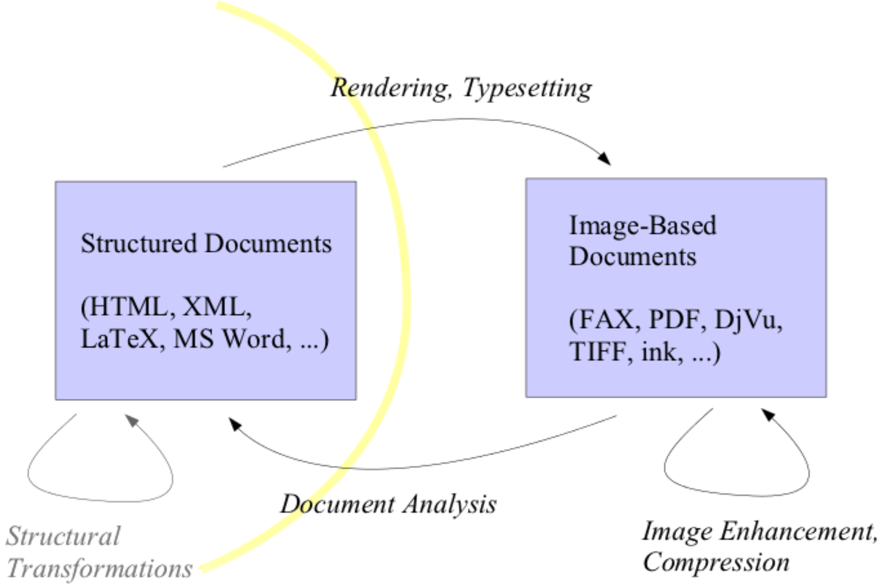

Rendering, Typesetting
Structured Documents
(HTML, XML, LaTeX, MS Word, ...)
Image-Based Documents
(FAX, PDF, DjVu, TIFF, ink, ...)
Document Analysis
Structural Transformations
Image Enhancement, Compression

Figure 1: The relationship between image-based and structured document formats. Typesetting generates images from structured document formats. Document analysis attempts to recover structured documents from image-based documents. Additionally, both structured documents and image-based documents can be transformed within their own types.

Traditionally, document analysis (or, more precisely, document image analysis, as the term "document analysis" is sometimes also applied to the analysis of the textual content of documents) is concerned with the conversion of documents in paper form into electronic formats[1]. Usually, paper documents are scanned, and the resulting document images are processed via document layout analysis and optical character recognition (OCR). This transforms them into a structured electronic format similar to that generated by word processors and other electronic authoring tools.

First, we should note, however, that the conversion of paper documents into electronic formats is still far from complete. There are large numbers of documents and books still in paper format that need to be converted. Furthermore, many documents are still published in paper format for a variety of reasons. For example, book publishers have been reluctant to publish in electronic

[1] We will not even attempt a survey of the extensive literature on document analysis in this paper; [10, 11, 2, 1] give an idea of the flavor of the field and [5] gives extensive references to different techniques that have been used in document imaging and document analysis

formats because of concerns about the potential for unauthorized copying and distribution. Legal documents and bill presentment still often take place in paper formats for legal reasons, concerns about privacy, and user reluctance.

Many cultures and languages also still use paper-based communications widely. Paper is also still a desirable medium from a forensic point of view: while modern electronic publishing methods have made it easy to create high-quality copies of documents and manipulate them, modern forensics has made it possible to detect even sophisticated forgeries and alterations of paper documents. And paper has desirable archival properties, proven to survive in some cases thousands of years, not yet matched or proven for other kinds of storage media.

Probably as a result of all these factors, as well as the ease with which paper-based documents can be created using printers, the use of paper is actually increasing [9].

Nevertheless, the common assumption in the industry and among computer scientists seems to be that paper-based documents are legacy documents and will be gradually replaced by the web, on-line forms, on-line publishing, and electronic handheld readers. Extensive work is being carried out on creating a wealth of standards for semantic, high-level representations of documents, standards like XML, DocBook, MathML, and SVG. Hardcopy documents are viewed as legacy documents and temporary intermediates, perhaps created for reading, and assumed to disappear over time.

With the extensive use of electronic content creation and distribution in electronic formats, therefore the question arises naturally whether there is still much need for document analysis, or whether document analysis will sink into obscurity after most paper-based documents have been converted into electronic format. In this paper, we examine that question. We will be arguing that:

- Paper is not going away for decades to come: a lot of engineering still needs to be done in order to make electronic alternatives competitive, both in terms of formats and in terms of electronic reading devices.
- Image-based representations and computations will become more important, not less important, in the future, even for documents that never get rendered in paper form.
- Hardcopy documents should be scanned, but they should be kept on-line in image-based form. OCR results should be viewed as an “annotation” of the document image, not as the definitive representation.

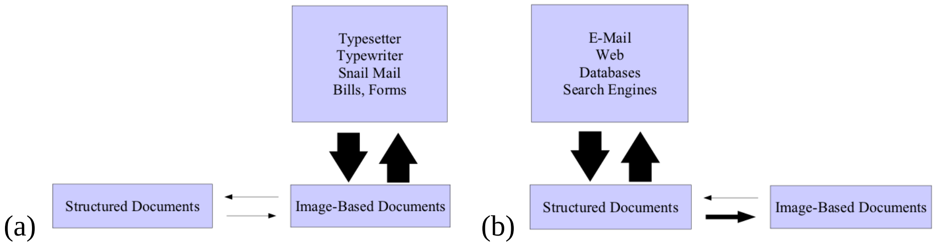


Figure 2: Traditionally, most documents were exchanged in image-based formats: paper, FAX, and others (a). The common view is that the future of document imaging will look as in (b), with documents created and exchanged in electronic format and occasionally printed out.

Documents and Document Analysis

The term “document” is used with many different meanings, both colloquially and in computer science, but generally, it refers to a substantial amount of human-readable information. Traditionally, paper-based documents consist of collections of printed pages with mixtures of text and graphics. Electronic documents come in many different kinds of representations. Two main categories of electronic documents we want to distinguish are *structured electronic documents* and *image-based electronic documents*.

Structured electronic documents are exemplified by formats like HTML, XHTML, XML, LaTeX, and Microsoft Word’s .doc format. They are used by office suites, web browsers, presentation packages, on-line forms processing systems, and many other software packages. Structured electronic documents usually contain an encoding of text in reading order using simple codes to represent individual characters or glyphs. The logical functions of different portions, like heading, page numbers, title, authors, sections, of the text are contained in annotations or “markup”. Structured electronic documents also contain annotations about the appearance of text: fonts, font styles, font size, colors, precise positioning. And they may contain drawings and digital versions of photographic images.

Structured electronic documents usually have well-defined syntax and semantics. This facilitates electronic processing and transformations of documents, but it also limits their representational capabilities. For example, the XHTML format is a well-defined structured document format for electronic publishing, but it lacks syntax and semantics for the representation of mathematics or images; without additional standards like MathML and SVG, XHTML documents cannot represent mathematical formulas or vector graphics.

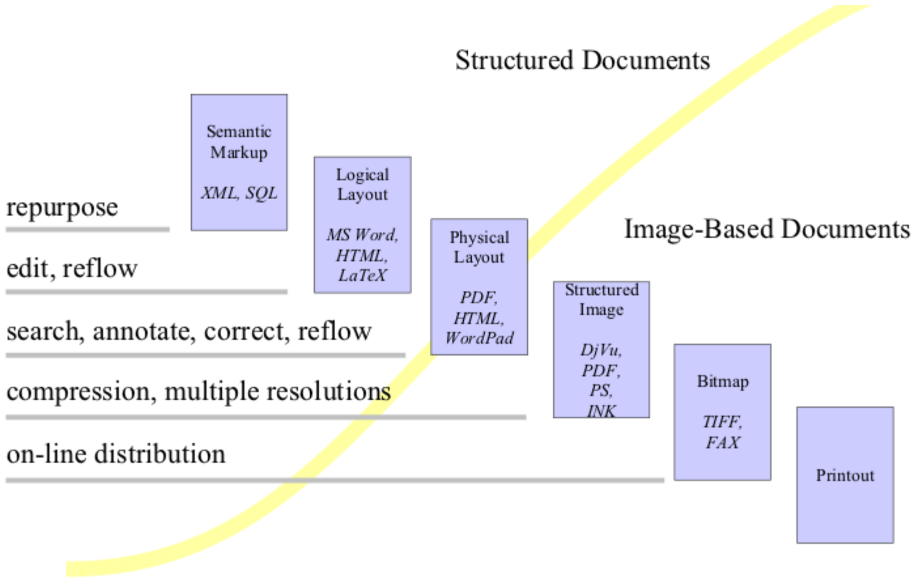


Figure 3: There is a wide range of possible electronic representations of documents, from representations close to images to representations of the text, layout, and content of a document.

Image-based electronic documents, in contrast, are a pixel-by-pixel representation of a document. They contain little or no information about reading order or the logical function of individual parts of the image. Sometimes, they are annotated with some textual meta information to facilitate document retrieval, but the image itself is the ground truth that the user sees and interacts with. Image-based electronic documents are created, for example, by scanning paper-based documents, but they are also created (temporarily or permanently) by printer drivers and window systems. Examples of image-based electronic documents are the TIFF graphics format, the DjVu token-based image compression format, various digital representations of "ink" for pen-based computing, and the PDF format (although the PDF format increasingly attempts to bridge the gap between image-based and structured electronic document formats by permitting the incorporation of structural information into an otherwise image-based format).

Image-based electronic documents have little or no syntax or semantics defined for them other than how the pixels themselves are compressed and represented. On the other hand, this lack of syntax and semantics also means that they can represent nearly arbitrary documents: a TIFF image of a page can contain text, graphics, mathematical notation, chemical formulas, drawings, and photographs. The only constraints imposed by image-based formats are constraints on resolution and color depth.

The division between structured document formats and image-based document formats is not quite as strict as described above. Actually, there is a hierarchy of formats and capabilities, illustrated in Figure 3. The more structure is present in the representation of a document, the more capabilities it supports. For example, simple bitmap formats permit on-line distribution, but little more. Formats like DjVu, image-based PDF files, Postscript, SVG, and ink permit rendering at multiple resolutions and yield additional compression. Font-based PDF documents, HTML, and RTF permit search, annotations, small corrections, and sometimes reflowing (adaptation to different output sizes). Document formats with logical layout information, like Microsoft Word, HTML, and LaTeX, permit editing and reflowing. And fully semantically marked up document types like XML or structured data can be manipulated and repurposed automatically (e.g., report writing, mail merging, etc.).

Traditionally, document analysis is viewed as a transformation of paper documents into structured document formats, but it can actually be a transformation of any kind of image-based document formats into structured document formats. In fact, as such, it is the reverse of document typesetting or rendering (this is analogous to how computer vision is often described as "inverse optics").

These different representations also correspond to different steps in OCR and document analysis systems. While a full OCR system is usually thought of as something that goes directly from a bitmap, it is actually a complex, multi-step process of transforming the raw bitmap image into a logically marked up text document (for a classic description of such a system, see [10]).

- The first step is to separate text from non-text in a document image, and to identify individual characters. The output of this operation can then be represented and compressed, yielding a highly compressed document formats like DjVu [6] and Digipaper [7].

- The collections of images and characters are then subjected to geometric analysis, character recognition, and language modeling. The output is a simple physical layout that can be used for PDF documents or simple transformations into HTML. However, the quality of the output is not comparable to that of documents originally authored in structured formats.

- In order to achieve a high-quality structured representation of the document, physical layout analysis needs to be followed by logical layout analysis, which assigns meaning to different parts of the input text. For example, it identifies section headings, titles, page

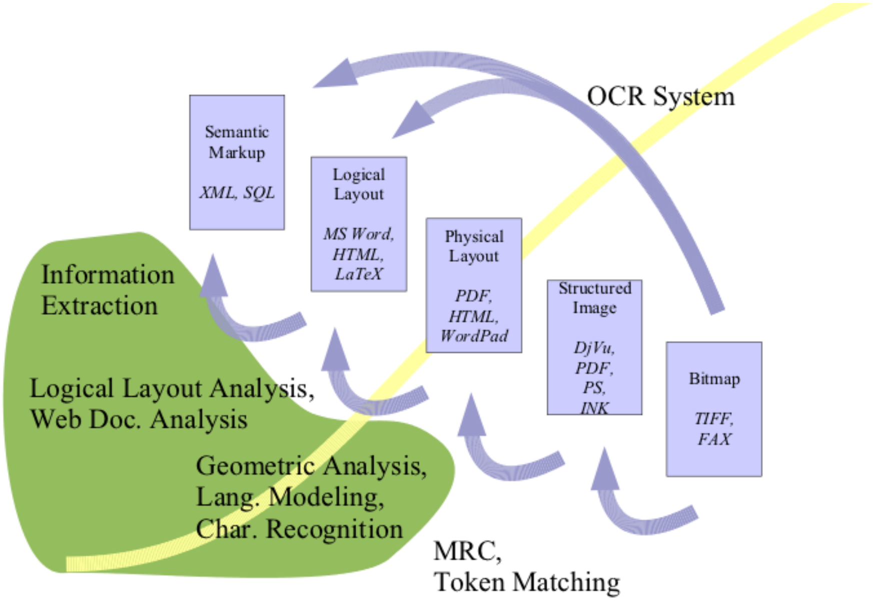


Figure 4: Document analysis proceeds in a series of steps, working up from image-based representations to high-level representations of content and layout.

headers and footers, captions, exact text flow, etc.

- Additional semantic information can be extracted using text analysis, content analysis, and information extraction techniques. This is a particularly active area of research because it applies to a wide range of electronic documents.

Electronic Readers

One of the key assumptions behind a migration to a world of electronic documents is that paper can be replaced. One of the key uses of paper is for reading. Therefore, one of the key questions behind a move to electronic documents is the question of whether we can replace paper for the purposes of reading. This actually does not include just the act of reading itself, but also locating, handling, annotating, and distributing paper.

Questions like these were looked into in Sellen and Harper's book "The Myth of the Paperless Office" [9]. Sellen and Harper look in detail at the *affordances* of paper–the things that one can do with paper easily. Rather than reiterating their arguments, let us ask what kind of high-quality electronic reader hardware we might imagine:

- high-quality paper-like display
- price $\$100$
- weight 250g (0.5 pounds)
- communications: IR, Bluetooth, WiFi, UMTS/3G
- format: A4-A6, constrained by user preference
- storage: 4G of flash memory
- battery > 24h battery life
- DRM: no DRM hassles

In fact, such a device is not that far off: the Sony Librie is a fairly lightweight device with a high-quality paper-like display. Its battery life and connectivity are somewhat more limited than our ideal device, and the device is hampered by digital rights management (DRM).

If we assume we have such an inexpensive and high-quality reading device, would it be competition for paper?

While such a device is attractive for many applications, it still has significant limitations compared to paper. We would not hesitate to read a printed memo, newspaper, magazine, or soft cover book at the beach, the pool, in the bathroom, and outdoors cafe, or a dining hall. While not completely impervious to moisture, spilled coffee, or salt spray, printouts, newspaper, magazines, and soft-cover books degrade predictable under the elements and generally remain readable. We can take a paper document on a long flight and not worry about the batteries running out. An electronic reader costing $100 may be inexpensive, but it is still far more expensive than a few printed pages, a newspaper, a magazine, or a softcover book–we don't have to worry about leaving, losing, or giving away any of the paper documents. We can annotate any of these using a standard pen.

If we work with multiple paper documents, we can spread them out. Yet, an electronic reader only displays one page of one document at a time; if we want to display multiple ones, we need to buy multiple devices and manage and synchronize all of them.

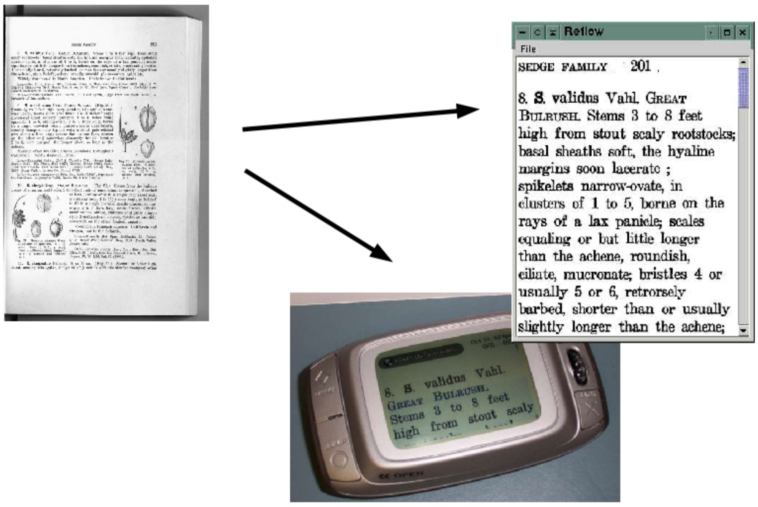
There are also legal and copyright problems of using such a device for reading. With a book, it is clear that I own and control the content: I can lend the book, I can sell it, and I can be certain that the ink does not disappear from the pages after a couple of years. With a digital reading device, none of those are obvious: the combination of digital rights management and electronic reader hardware may make it impossible to lend or sell individual books, or the license may restrict use to a single named individual. And the content itself may become inaccessible after a limited period (for example, books on the Librie are available only for a limited amount of time).

If we would like to have a digital reading device that can truly compete with paper, some significant additional technical advances are still needed:

- Eliminate the need for batteries or recharging completely–perhaps the device could be power with solar cells, like some calculators.
- The weight should be less than 50g and the thickness less than 5mm, and the device should be flexible and ideally even roll up. It should be sealed and water-resistant.
- The cost should be around $10/display, so that it is easy to buy multiple devices that can be spread out across a table, given away, etc. A low price also makes loss and damage less of a concern.
- The device needs a fast, robust, and nearly invisible operating system. Networking should be transparent and permit transparent synchronization across devices and with servers.
- A physical user interface that supports annotations, search, and document exchange with simple, intuitive gestures.

While the technologies making such a device possible are in development, it will likely take a considerable amount of time until the cost of such a device can come down to the desired level.

So, we see that even a fairly high-quality reading device that can be built with current technologies within a few years still has some limitations. However, even with its limitations, they still have some strengths that make them very useful in some applications, even if they cannot compete with paper in other areas. Some of these strengths are the ability to search through large amounts of information, update content, represent large amounts of information in a small device, and display information that is already in electronic form. This makes current electronic readers particularly suitable for applications like reference works (frequently



SEDGE FAMILY 201.

8. S. validus Vahl. Great Bulrush. Stems 3 to 8 feet high from stout scaly rootstocks; basal sheaths soft, the hyaline margins soon lacerate ; spikelets narrow-ovate, in clusters of 1 to 5, borne on the rays of a lax panicle; scales equaling or but little longer than the achene, roundish, ciliate, mucronate; bristles 4 or usually 5 or 6, retrorsely barbed, shorter than or usually slightly longer than the achene;

8. S. validus Vahl. Great Bulrush. Stems 3 to 8 feet high from stout scaly

Figure 5: An example of a paper-to-PDA conversion. A scanned full-page image is shown on the left. After processing, it can be displayed in a resizable window or on a cellular telephone.

updated, searchable, usually only small portions read at a time), school textbooks (in paper form, they are outdated quickly and heavy), web and email texts (frequent updates and already in easy accessible electronic formats), and some forms (mail order, government forms, taxes).

On the other hand, the limitations of current or near-future electronic reading hardware makes other kinds of reading materials still more suitable to reading on paper. For example, location bound documents or forms, like guides, sign-up forms, tourist information, human resource forms, flyers, handouts, etc., are more easily distributed in paper form and filled in using a regular pen: doing so can be accomplished without any kind of concerns about networking and format compatibility, the cost of distributing the physical artefact is low, and users know how to interact with the documents. Likewise, for beach reading, travel reading, and travel documents, the predictable availability, low cost, and robustness of paper still win out. Also, for business, technical, and scientific reading, the ability to annotate and distribute paper documents, as well as the high resolution and the easy need to interact with multiple documents at once are important advantages of paper.

Opportunities for Document Analysis in Electronic Reading Devices

While document analysis cannot directly address the hardware limitations of current electronic reading devices, it can help in some other areas. Document image analysis can

- simplify the conversion of structured documents into formats suitable for display on electronic reading devices
- improve capture and conversion of paper documents for reading on electronic reading devices
- create long-lived, well-documented open document formats
- transfer some of the affordances of paper documents to electronic reading devices, including annotations and intuitive sharing mechanisms

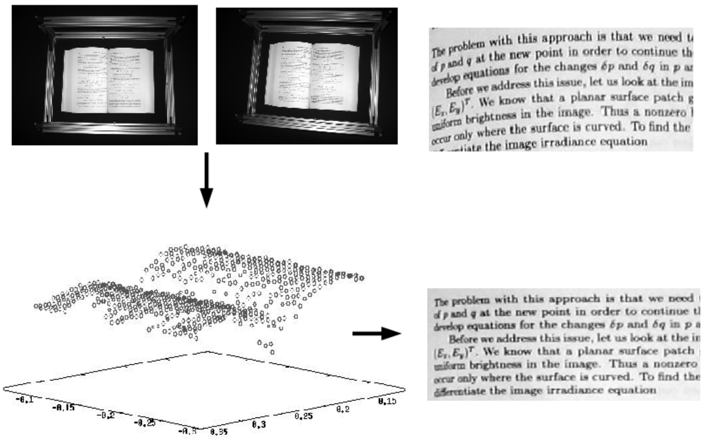
We will discuss such approaches in more detail below.

Conversion of Paper to Electronic Book Formats

One of the limitations of electronic books is the lack of availability of some types of reading materials. While content like web pages and electronic mail can be easily read on electronic readers, other content, like books, scientific papers, and business memos are either not available in electronic form at all, or are available in electronic formats like PDF that are designed for printing on paper but of limited suitability for display on small-screen reading devices. Many current electronic formats also cannot represent complex information like mathematical or chemical formulas well.

One approach to converting papers to electronic book formats has been described by [3]. The idea behind the approach is to start with a scanned or digital image of a document and perform physical layout analysis (see above). The result of physical layout analysis is a collection of text columns, images, and word bounding boxes, together with an approximate reading order.

The result of physical layout analysis is used to generate of the original document consisting of a mixture of text word images and non-word images in approximate reading order. Unlike the original scanned image, this collection is *reflowable*, that is, it can adapt automatically to



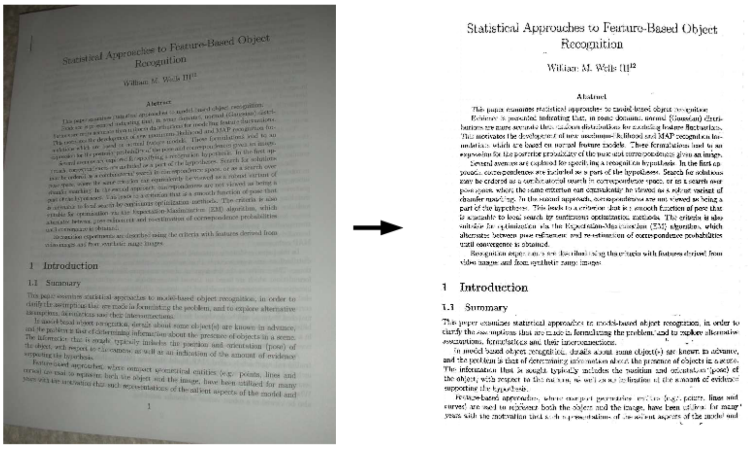
The problem with this approach is that we need t
of p and q at the new point in order to continue th
develop equations for the changes δp and δq in p ar
Before we address this issue, let us look at the im
$(E_x, E_y)^T$. We know that a planar surface patch g
uniform brightness in the image. Thus a nonzero l
occur only where the surface is curved. To find the
[illegible] the image irradiance equation

The problem with this approach is that we need
of p and q at the new point in order to continue t
develop equations for the changes δp and δq in p a
Before we address this issue, let us look at the in
$(E_x, E_y)^T$. We know that a planar surface patch
uniform brightness in the image. Thus a nonzero
occur only where the surface is curved. To find the
differentiate the image irradiance equation

Figure 6: An illustration of camera-based document capture using stereo vision methods. Two images of a document (book) are captured with digital cameras from slightly different viewpoints. By applying stereo vision and surface interpolation techniques, a 3D surface can be reconstructed. This can be used for dewarping the distorted image into a planar representation of the page image.

screens of different sizes. Experience with this system shows that complex logical layout analysis, like the identification of floats or page headers/footers, is not required in order to obtain a readable rendition of the text. This collection of images can ultimately be converted into one of a number of exiting electronic book formats, including Plucker, OpenEbook (OEB), XHTML, or reflowable PDF for display on existing handheld devices. Alternatively, it is possible to maintain the original page images (for example, in TIFF or DjVu format) and represent the necessary layout and reflow information as a separate annotation; such an annotation tends to be small compared to even a highly compressed representation of the bitmap image.

Such reflowable representations of scanned documents can be viewed on a variety of displays, including smaller laptop screens, PDAs, or even cellular phones (Figure 5). The representation is compact and simple, in general no larger than a TIFF or PDF representation of the content, and, appropriately represented, about the size of a compressed representation using token-based compression [6, 7]. Regardless of how it is represented, the representation contains the complete and exact original document image. Such a representation is also universal: it can represent any kind of document, even ones containing math, chemical formulas, and diagrams; no new markup is required to represent arbitrary content. Furthermore, compactness, simplicity, universality, and preservation of the original document image make it a good candidate for



Statistical Approaches to Feature-Based Object Recognition

William M. Wells III[12]

Abstract

This paper examines statistical approaches to model-based object recognition. Evidence is presented indicating that, in some domains, normal (Gaussian) distributions are more accurate than uniform distributions for modeling feature fluctuations. This motivates the development of new maximum-likelihood and MAP recognition formulations which are based on normal feature models. These formulations lead to an expression for the posterior probability of the pose and correspondences given an image.

Several avenues are explored for specifying a recognition hypothesis. In the first approach, correspondences are included as a part of the hypothesis. Search for solutions may be ordered as a combinatorial search in correspondence space, or as a search over pose space, where the same criterion can equivalently be viewed as a robust variant of chamfer matching. In the second approach, correspondences are not viewed as being a part of the hypothesis. This leads to a criterion that is a smooth function of pose that is amenable to local search by continuous optimization methods. The criteria is also suitable for optimization via the Expectation-Maximization (EM) algorithm, which alternates between pose refinement and re-estimation of correspondence probabilities until convergence is obtained.

Recognition experiments are described using the criteria with features derived from video images and from synthetic range images.

1 Introduction

1.1 Summary

This paper examines statistical approaches to model-based object recognition, in order to clarify the assumptions that are made in formulating the problem, and to explore alternative assumptions, formulations and their interconnections.

In model-based object recognition, details about some object(s) are known in advance, and the problem is that of determining information about the presence of objects in a scene. The information that is sought typically includes the position and orientation (pose) of the object, with respect to the camera, as well as an indication of the amount of evidence supporting the hypothesis.

Feature-based approaches, where compact geometrical entities (e.g., points, lines and curves) are used to represent both the object and the image, have been utilized for many years with the motivation that such representations of the salient aspects of the model and

Figure 7: A simple dewarping technique for single frame camera-based capture. A novel text line finding technique, capable of finding non-parallel text lines in perspectively distorted images, is used to determine the 3D distortion of the captured page image. A subsequent dewarping step reconstructs a planar, undistorted version of the image.

long-term archival storage.

Simplifying Document Capture

Another significant problem with electronic documents is that capturing them is difficult compared with paper-based document distribution. A piece of paper, a memo, or a bound brochure can be distributed quickly and easily, anywhere, with no special hardware. No comparable distribution mechanism exists yet for electronic documents. Distribution as “digital objects” via infrared or Bluetooth was intended to permit a similarly simple distribution, but such mechanisms are slow and fraught with incompatibilities.

Capturing document images is currently primarily carried out using scanners. High-resolution desktop scanners are available inexpensively but take of the order of 30-60 seconds per page in scanning time, plus a significant amount of setup time. High-speed optical document capture is currently carried out using bulky and expensive scanners. There are some small, portable document scanners that permit a page at a time to be scanned by moving a handheld wand across the page, but their quality is limited and the scanning process with them is still slow and cumbersome.

Because of such limitations, *camera-based document capture* is of considerable interest [12, 4, 8, 14]. With camera-based capture, we can "pick up" a page simply by pointing a camera at it. Experience with small, lightweight cameras like the Sony DSC-T1 shows camera-based capture to be convenient and quick. Cameras like the DSC-T1 also offer a large, readable screen that permits a quick and immediate review of captured document images.

However, camera-based document capture still faces significant obstacles: document images captured with a camera differ considerably from those captured on a flatbed scanner in that they show both large geometric distortions and non-uniform lighting. One of the primary challenges of camera-based document capture is to correct these deviations from an ideal captured image. A number of approaches for camera-based document capture have been described in the literature. As mentioned above, common ones rely on structured light techniques or shape-from-shading (also applicable for correcting scanner-based distortions).

In our own work [13], we have pursued camera-based document capture using uncalibrated stereo and motion vision techniques. Our approach permits document images to be captured quickly, using either multiple shots with an unmodified digital camera, or a single shot with a digital camera set up for stereo photography (a simple and inexpensive modification to an existing camera). Stereo vision is used to derive a 3D model of the surface of a captured document. This model is usually sparse and noisy, and we apply robust surface interpolation techniques in order to recover an approximation to the true 3D shape of the page surface. From the 3D shape of the page surface, we can compute a dewarping transformation that transforms the image back into a flattened image, comparable to an image captured with a flatbed scanner. The approach is shown in Figure 6.

In a second approach, we apply a novel line finding technique to perspectively distorted page images. This permits page dewarping of page images from a single image. This is illustrated in Figure 7.

Camera-based document capture solves a number of problems with the distribution of documents for handheld readers: it allows casual "grabbing" of a document as easily as pointing a camera at it. The approach is similar to picking up a page, with some differences. An advantage is that the physical document can be left behind and that capture and storage is under a recipient's control. A disadvantage is that the time required for capturing a document is proportional to the number of pages.

Hardware for stereo-based capture need not be significantly more expensive than hardware for traditional digital camera capture; we are currently constructing prototypes. On-going work in our lab is to improve sparse surface interpolation further and to achieve better dewarping from single document images.

Adding Paper Capabilities to Electronic Documents

Let us mention two more areas where document analysis and related techniques can contribute to improving the usability of electronic document readers.

The first area is that of making documents easier to annotate. Structured documents can be annotated using “electronic ink”, a compressed, vector-based representation of handwriting and drawing. Microsoft has taken that approach in Microsoft Windows, Microsoft Office, and other applications. Electronic ink integrates well with structured representations of document images. Annotations can also be represented in an image-based framework as separate layers; this kind of representation is supported by formats like TIFF.

A second important capability we have already mentioned is the ability to work with multiple documents on multiple displays simultaneously, an ability that is quite natural with paper-based documents. This raises significant human-computer interaction (HCI) issues, as well as networking and security problems. We have explored user interfaces based on simple spatial gestures for transferring documents between devices, but significant additional work is needed.

In summary, in this section we have seen that we can improve content availability for electronic readers, as well as achieve archival quality, through developing reflowable image formats. Camera-based document capture permits us toe acquire and distribute document images easily using handheld cameras. Developing support for annotations through layered image-based representations permits additional capabilities of paper documents to be carried over to digital documents. Overall, we see that document analysis and related techniques are important to making electronic documents and electronic document readers more paper-like.

Paper/Electronic Integration

Even though techniques like those described in the previous section may eventually permit fully electronic handling of documents, the advantages of paper mean that it will likely continue to be used for a long time. However, by working on the integration of paper and electronic documents, we can potentially achieve the best of both worlds.

Of course, on-line distribution of documents followed by printing is already a common way of combining electronic documents with paper-based documents. What is still lacking is simple and effective means of round-trip integration of paper and on-line versions of documents. For example, once a document is printed, we would like to be able to annotate it, make corrections to it, and then retrieve the corresponding electronic document and reintegrate the changes with the on-line version of the document. Although we will not say much more about it in this paper, we note that one particularly important example of round-trip paper handling is forms processing: forms are designed on-line, filled in with pen and paper, then scanned, and the data contained in them is extracted and transformed into electronic formats. Significant amounts of work have been done on this problem in the past, although the techniques developed up to this point have not found much application in practice.

One significant obstacle to the use of round-trip techniques is probably the fact that document capture is cumbersome. We have already mentioned handheld camera-based document capture above. However, for desktop usage, we are currently developing a different set of camera-based capture techniques that permits fast, oblivious document capture on the desktop.

Current commercial camera-based document capture solutions still require significant user interaction and take up valuable desktop real estate. We have developed a set of techniques relying on high-resolution cameras that can monitor and capture the desktop in real time. Through a combination of careful camera placement and optical elements, we achieve an unobtrusive integration of the camera into the user's physical desktop. Finally, through motion analysis and gesture recognition techniques, capture becomes oblivious and automatic. The long-term goal of this work is to automatically capture and archive everything that the user sees and works with.

For annotations, there are several important cases of round-trip handling we can distinguish:

- retrieving the electronic version of a document from the scanning or camera-based capture of a paper document
- separating printed matter from handwritten annotations without knowledge of the original electronic document
- detecting and separating annotations assuming the exact electronic version of the document is known
- comparing electronic documents and scanned documents even if they are different versions (with layout changes, reflow, etc.)

- comparing two scanned documents, possibly allowing for layout changes

Several of these cases are covered by classic document analysis papers. We are currently developing specifically image-based document comparison methods with applications to the analysis of historical documents, as well as to forms handling.

Image-Based Personal Computing

We started this paper with the observation that there are different levels of representations of documents, which we coarsely categorized into structured representations and image-based representations. In the real world, humans almost exclusively interact with image-based representations of documents: typeset and rendered documents, newspapers, magazines, books, etc. Yet, on-line, most documents are still represented in structured form.

Historically, the reason for this situation is probably that until recently, capture, storage, compression, and transmission of images was costly and required more storage and CPU power than was generally available. Furthermore, there was a lack of good methods for manipulating and transforming image-based documents. Finally, there was a bias of computer scientists and software developers towards formal languages, syntactic correctness, and precise data models. As a result, structured documents were effectively the only choice users effectively had for representing documents.

It is unclear whether this reliance on structured representations is desirable. Structured representations, in fact, place significant burdens, both on software authors and on users. Modern WYSIWYG interfaces are, in fact, a carefully maintained illusion: while the underlying representation is structured, the GUI does not usually show that structured representation.

For example, when cutting and pasting between two applications, the user perceives the operation as selecting a region on-screen containing some text, which is then somehow inserted into another application. However, the actual operation is considerably more complicated: the mouse gesture used to start the cutting operation is transmitted to the underlying application, which tries to match up the screen coordinates with the structured representation; the parts of the structured representation that correspond to the on-screen display are then transmitted to the target application. In this illusion, many things can go wrong: numerous parts of a user interface do not, in fact, support cutting and pasting at all. For example, text labels and text in dialog boxes usually cannot participate in cut-and-paste operations.

PDF documents are particularly counterintuitive from a user's point of view and illustrate the mismatch between structured and image-based representations well. PDF documents can encapsulate both structured and image-based representations. Some PDF documents are fully structured representations and support both reflowing and cut-and-paste operations, via mechanisms similar to those describe above. Other PDF documents, visually indistinguishable from an end user's point of view, are fully image based; such documents do not permit cutting and pasting at all. Yet other PDF documents display an image-based representation but back it with an errorful OCR representation of the same document (such documents are generated by Adobe Acrobat). Such documents appear to support cutting and pasting, but the user discovers when attempting to paste the text that the text often contains many errors compared to what appears to be displayed on-screen.

While, in the past, the reliance on structured representations was forced by both available hardware and software, the situation has changed dramatically since. Image capture has become much simpler, bandwidth, storage, and compute cycles have become comparatively cheap. Furthermore, there are many new sources of image-based documents, like ink from Tablet PC computers and PDAs. Standard desktop PCs have become fast enough for complex image manipulations, layout analysis and character recognition, and image matching and content-based retrieval. There have been great advances in pattern recognition and document image analysis techniques. And a new generation of computer scientists and software developers are looking at adaptive software systems and integrating statistical machine learning into day-to-day computer usage.

We can therefore imagine computing environments in which both computers and users share a common view of documents: an image-based view. In such a view, the discrepancies that exist with structured representations disappear. For example, a cut-and-paste operation is not implemented in terms of some complex behind-the-scenes protocol, but relies on a direct interpretation of the image as it is displayed on-screen. Such an implementation of cut-and-paste operates universally, irrespective of the particular implementation or representation of content that an application is using[2].

A comparison of capabilities of structured and image-based document representations is shown in Table 1. We note a number of advantages of image-based representations over structured representations of documents: more intuitive behavior, ability to represent arbitrary content, simple format conversions, lower up-front investment in markup, and improved privacy and security. On the other hand, structured representations currently still have some advantages: we can change their presentation easily through style files, they potentially allow more complex

[2] This particular capability has already been demonstrated by some shareware add-ons to Windows.

Structured Representations	**Image-Based Representations**
Advantages of Image-Based Representations	
Representation disagrees with what the user sees/expects.	Representation corresponds to what the user sees/expects.
Can represent only content for which markup has been defined	Can represent any kind of content.
Difficult to convert correctly between different formats.	Easy to convert between different formats.
Requires lots of up-front work to create correct markup.	Once it looks right on-screen, you are done.
By design, contains hidden metadata–security and privacy issues.	The data you see is all the data there is
Advantages of Structured Representations	
Presentation easy to change through style files.	Presentation difficult to change.
Allows complex searches.	Allows keyword searches.
Allows data linking/reuse (spreadsheets, mail-merge).	No tools for data linking/reuse available yet.

Table 1: Advantages and disadvantages of structured and image-based document representations compared. See the text for a discussion.

searches, and they simplify data linking and reuse.

However, these remaining limitations of image-based document representations are surmountable. For example, changing the presentation of an image-based document can be carried out by applying OCR/layout analysis and recovering a temporary structured representation. Then, the formatting of that temporary structured representation is altered, and the document is finally rendered in image form again. Such an approach would be difficult to apply right now, because current OCR and layout analysis systems are not yet sufficiently robust to rely on creating a correct structured representation of an image based document that permits reformatting, but with improvements in both OCR/layout analysis dependability and performance, such an approach will become quite simple and natural in the future.

More generally, an image-based approach to personal computing, as we are suggesting above, assumes that components capable of transforming unstructured input into structured input on

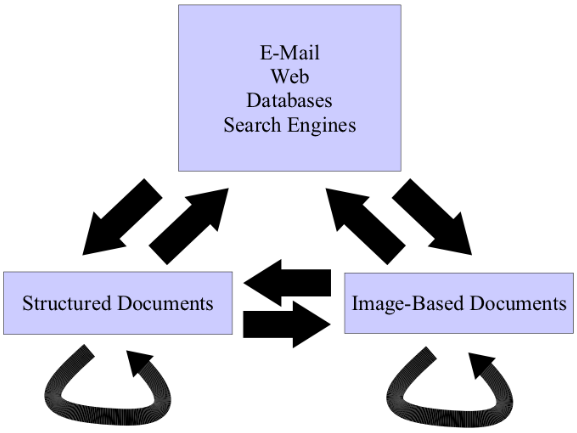


Figure 8: We believe that in future personal computer systems, rather than disappearing, image-based document types will play a major role along-side structured document representations, and that all aspects of conversion, transformation, and communication of both structured and image-based documents will have to be addressed.

the fly are available. While a fully image-based approach still requires considerable amounts of research and systems development, even existing systems are starting to incorporate such functionality already. For example, both Microsoft Word and OpenOffice include "auto stylist" functionality, in which a user indicates the formatting of text (bullet items, spacing, alignment, etc.) by approximating the intended appearance in plain text, without markup, and the stylist infers from the user's input what the intended high-level layout is supposed to be. Similarly, Wiki web editors are becoming increasingly popular, and they, too, infer high-level formatting instructions from the actual spatial arrangement and appearance of input text provided by the user.

Overall, we believe that image-based representations will become increasingly important in all aspects of computing (Figure 8). End users will pick and choose the right representation according to task. In order to support this, operating systems and toolkits will need to have easy-to-use support for common transformation and imaging tasks, like

- built-in reliable, robust OCR and layout analysis functionality
- content-based and appearance-based document image retrieval
- better abstractions and user interfaces for image processing tasks

- support for image data types as extensive as for XML and ASCII

Conclusions

In this paper, we have looked at a number of aspects of electronic documents, in particular from the point of view of the distinction between structured documents and image-based documents. We have seen that document analysis is a collection of technologies necessary for transforming image-based documents into structured documents. The degree to which document analysis has a future in the era of electronic documents depends on the degree to which we believe that image-based document representations are going to be important in the future.

We have argued that traditional uses of document analysis, namely the conversion of paper-based documents into structured formats, will continue to be crucially important in the future because paper itself shows no signs of disappearing. And some key technologies necessary for replacing paper, namely electronic document readers, themselves require document analysis for purposes such as capture and conversion of materials into formats suitable for such readers.

Our long-term vision is that image-based document representations will become increasingly important in personal computing. While, in the past, they have only been used as temporary intermediates between, say, document scanning and OCR software, increased compute power and bandwidth, as well as improved document analysis technologies and software, will make it increasingly feasible to use image-based representations as the primary representations of documents on computers. We believe that such representations will become widespread because they are far more intuitive and flexible than structured representations.

References

[1] Henry S. Baird. Model-directed document image analysis. In *Proceedings of the DOD-sponsored Symposium on Document Image Understanding Technology (SDIUT 1999)*, Annapolis, MD, April 1999. Invited published talk.

[2] Henry S. Baird and George Nagy. A self-correcting 100-font classifier. In *Proceedings of the IS&T/SPIE Symposium on Electronic Imaging: Science & Technology*, San Jose, CA, February 1999.

[3] T. Breuel, W. Janssen, K. Popat, and H. Baird. Paper to PDA. In *Proc. 16th Int. Conf. on Pattern Recognition (ICPR)*, 2002.

[4] Michael S. Brown and W. Brent Seales. Document restoration using 3D shape: A general deskewing algorithm for arbitrarily warped documents. In *International Conference on Computer Vision (ICCV01)*, volume 2, pages 367–374, July 2001.

[5] R. Cattoni, T. Coianiz, S. Messelodi, and C. M. Modena. Geometric layout analysis techniques for document image understanding: a review. Technical report, IRST, Trento, Italy, 1998.

[6] Patrick Haffner, Leon Bottou, Paul Howard, and Yann Le Cun. Djvu : Analyzing and compressing scanned documents for internet distribution. In *Proceedings of the International Conference on Document Analysis and Recognition*, pages 625–628, 1999.

[7] Huttenlocher D., Felzenszwalb P., and Rucklidge W. Digipaper: a versatile color document image representation. *Proceedings of 6th International Conference on Image Processing (ICIP'99)*, pages 219–23 vol.1, 1999.

[8] Maurizio Pilu. Undoing page curl distortion using applicable surfaces. In *Computer Vision and Pattern Recognition Conference*, pages 67–72, December 2001.

[9] A. Sellen and R. Harper. *The Myth of the Paperless Office*. MIT Press, 2001.

[10] S. Srihari. Document image understanding. In *Proceedings of 1986 ACM Fall joint computer conference*, pages 87–96, 1986.

[11] Y. Y. Tang, C. Y. Suen, C. D. Yan, and M. Cheriet. Document analysis and understanding: A brief survey. In *Proc. Int. Conf. on Document Analysis and Recognition*, pages 17–31, 1991.

[12] M. J. Taylor, A. Zappala, W. M. Newman, and C. R. Dance. Documents through cameras. In *Image and Vision Computing 17*, volume 11, pages 831–844, September 1999.

[13] A. Ulges, C. Lampert, and T. Breuel. Document capture using stereo vision. In *ACM Symposium on Document Engineering*, 2004.

[14] Zheng Zhang. Restoration of curved document images through 3D shape modeling. In *International Conference on Computer Vision and Pattern Recognition (CVPR2004)*, June 2004.