MEng Thesis Proposal
Adam Holt

SCAN YOUR LIFE:
Integrating OCR Into your Personal Haystack!

1. BACKGROUND

Optical flatbed scanners cost hundreds of dollars until 1997, when competition caused prices to collapse, often to below $50 bundled with last year's OCR software. This democratization of flatbeds dried up the market for handheld scanners, even forcing market grand-daddy HP to release its own $99 flatbed scanner. The only truly good reason left not to own a scanner in this era may be the valuable real estate on your (physical) desk.

America took notice: suddenly 25% of households with Internet access own a scanner [www.infotrends-rgi.com/press/1999092089476.html], versus only 8% that own a digital camera. These 10m scanner households represent roughly the same 10m that have added a separate computer line (of 100m total US households.) Microsoft took notice: their fall 1999 licensing of Xerox/Scansoft's OCR indicates that Windows98's successor will include not only voice but also document recognition.

For all the grief about OCR error rates, optical character recognition is a remarkably mature technology. Like the endless whining about digicam resolution, this supposed inadequacy is more likely a reflection of ergonomic and software integration shortcomings. So I wish to explore a technology that was first used for forms entry in 1959 -- and is so fast today that most off-the-shelf PC OCR users don't even notice the computation after the time-consuming physical scan. High-end Xerox digital copiers now convert this perception to reality, taking over the OCR job from your computer, embedding it in the copier.

Accuracy remains imperfect but is nothing to sneeze at: notoriously lo-res faxes (200dpi) are very often OCRable today. Remember, the best-skilled expensive human secretaries and stenographers make mistakes too. Hence dual-pass and triple-pass blind key data entry, used by competitive phone books and netlibrary.com's race to digitize literature, specifically labor-intensive redundant keyboarding done by Indians, Chinese and Philippinos. Of course OCR speeds of even 300 cps on an old Pentium are 100 times faster than trained manual keying speeds of 3 cps (average based on regimented, heads-down data entry tasks, according to the Association for Work Process Improvement.) In short, human-machine hybrids "each proofreading each other" are best today. Hypothetically of course, it may one day be possible for a computer to read marred text not only more cheaply but also more accurately than humans (voice recognition advocates have made similar claims.)

40 years of Moore's Law did more than improve an already successful application. Contemporary OCR software goes miles beyond its namesake; what was decontextualized character recognition ought be rechristened Optical CONTENT Recognition. Though much OCR remains close to its roots: mundane forms and records capture. But Boston-based

Xerox/Scansoft's TextBridge Pro 9.0, which Microsoft licensed for future products in November 1999, goes so far as to generate HTML that captures much paper document structure, starting with bolding, columns, tables. So why not begin unifying 2 of the biggest repositories in your life today (all your paper with all your efiles/email?)

Such "reality merge" would allow you to rapidly search all avenues of your life's info-artifacts, physical and virtual, minus Negroponte's Berlin Wall between atoms and bits. How many times have you been stuck, wondering "Where on earth was that illuminating article I read only just last week?" True, one big messy pile on the library floor may be no better than two (the Internet often being compared to a library with all the books dumped on the floor.) But now your own Haystack can impose rich internal structure around your files, one and all. Even simple keyword search is something profoundly missing from papers and books -- unless the author got carried away with his index. Is OCR another rich untapped technology emerging from gestation?

From Fiber To Fiber: legacy fiber [paper] use continues quite resiliently against the onslaught of fiber [optics]. Yet web pages are read on-screen more and more, and "save-a-tree" holds even more true with email. An experiment at the most populous campus in the USA (utexas.edu at Austin) very unexpectedly showed that people will not only read, but demand portable e-books, even with only 6000 titles. Likewise business and consumer billing is slowly but surely moving online across entire sectors of the economy. Of course, nothing may save us from the inundation of disposable "junk" fliers and handbills. Even if paper were to disappear, its aura will pervade forever in (1) artificial (but functional) paper such as eink.com and (2) bookshelves as decorative cultural/intellectual medallions (not to mention metaphors.)

Internet ideology forever turns on enhancing rights and democracy by making buried information more easily available to citizens. Yet past purchases of a book, or tuning to a broadcast radio or television channel, did not necessitate the acquisition of a copyright licence. Modern purchase, rental or even viewing of digital media, on the other hand, generally does. So OCR could be one of the only tools left with potential to empower individuals, to maintain control over their "collection" of personal effects -- in the face of rapidly increasing legal, contractual and technological attacks on "fair use" of downloaded material. In fact, personal digital imaging may be tantamount to a 1st Amendment speech right.

Historically it has been very difficult to achieve quality bulk-scanning of legacy documents for significantly less than $1 per page. Much of the cost arises from exception-handling: clearing paper jams, ever-changing fonts, page styles, colors schemes, etc. So despite home scanners' surging popularity, most are probably only used a couple times a month, and mainly for art projects. Certainly no harm there (!) but image management is only a secondary objective of this Thesis. The point: long-term scanner ergonomics are unclear -- basic vision software might allow digital cameras to double as a consumer scanners in the visible future. In fact, high-end consumer digital cameras (1600x1200 today) will soon surpass Group 3 fax (2200x1700 for 8.5x11 paper.) Indeed scanning and OCR cannot but spread like wildfire when tedious mechanical passes become unnecessary. For now these devices remain

pricey even to the elite, and the disposable cam atop your monitor has 640x480 resolution at best. But imagine: someday even your cursive scribbles may never be truly private again...

Viable business models for OCR software usage or bundling are equally unclear. For example, one could imagine imagine notarizing a photograph of one's actual bookshelf to be eligible for the discount "upgrade price" when viewing said titles electronically. IE. publishers will generally favor amortizing the cost of a much higher quality scan across all readers (who lose their valuable margin annotations!) Not incidentally, publishers can then look over your shoulder to acquire detailed user profiles.

2. PLAN

This thesis will develop a fully integrated scheme to, scan, OCR and archive documents into individuals' personal Haystack haylofts. Much like you can point Haystack at a tree of directories you want archived, you might like to "point" Haystack at a not-entirely organized collection of your papers, and recover much of its long-forgotten order. Haystack services will be developed to incorporate and usefully process the hundreds of interesting paper sheets that surround you, unnoticed.

Very early on the choice will be made between an auto-feed scanner and a flatbed. This scanner will be purchased or borrowed in January. Hedging for a (semi-automated) scanner is also possible; for example the quick $499 HP 6350Cse attempts to combine the best of both these scanning techniques. The $799 HP R80 is an all-in-one unit that adds to this faxing, color printing and color copying. Scan a stack of 20 sheets at once with these feeders, or use the flatbed for books with spines. Supporting low-quality 200dpi fax scans will also be considered in light of the rash of recent "free" fax-2-email services like (efax.com fax4free.com callwave.com jfax.com etc) that indeed offer you "no more paper-jams." For all its faults, this approach is based on an open standard (Group 3 int'l fax,) platform-independent across both phone networks and Internet, offering attractive worldwide ubiquity.

Unfortunately, open source OCR has not kept up as the bottom-heavy Linux world continues to lack rich applications. See http://documents.cfar.umd.edu/ocr/ and www.cfar.umd.edu/~kia/ocr-faq.html Specifically, both the two main Linux OCR projects are very forlorn, http://starship.python.net/crew/amk/ocr/ officially has "fallen into a coma" since mid 1999 and socr.org is untouched since November 1998. Worse, modern OCR applications tend to (1) be Windows-exclusive and (2) have GUIs that automatically pop up. We may have to tolerate Windows as an input device for now -- or possibly not (vividata.com uses Caere Omnipage's engine on UNIX but only outputs text for Linux.) Solving or mollifying (2) is a much higher priority concern that will be dealt with straight off. A likely solution may be interfacing to Caere's engine on Windows.

Archivists and digital archeologists are the ultimate packrats, never letting you throw anything away. Contemporary OCR doctrine goes yet further, advocating the wise preserving intermediary digital image scans, anticipating inevitable OCR algorithm improvements. Perhaps to later archive handwriting notes you left in the margin -- if only just

applied on demand and just-in-time to the most interesting parts. Yet images always entail bandwidth costs: a single 8.5x11 sheet's uncompressed scan varies from sub-floppy 400K (1bit 200dpi fax) to 400MB (24bit 1200dpi art), which is most of a CD! How far can we push Haystack's original "disk is free" philosophy? Everyone demands that disk and tape systems be cheap, available, and trusted, yet this remains crucially untrue in all but rare cases. While photoscans and faxes are becoming legally binding, "original" paper copies continue to be preserved at great expense, if not for legal then for archival assurances -- even as digitization buys you much better search accessibility.

Attempts will be made to preserve text substructures such as integrity of bulleted sections and chapters, all across page separations, and if possible integrated images. Certainly we will permit individuals to scan in mere snippet excerpts as well. Extracting as much paper doc structure as possible into HTML format is the most likely intermediary, if this most recent trend in 1999 OCR packages delivers on its promise. Though proprietary PDF or Word is also possible -- whatever format, raw or not, docs will likely be emailed to the user as MIME attachment (some digital copiers already offer this) for easy further processing by Haystack services.

That certain OCR packages (wonderfully) extract metadata such as physical addresses in your documents presents some complications however. How best to preserve such valuable Windows annotations, transfering them into Haystack? ScanSoft's OCR is integrating HP's JetSend protocol for rich inter-appliance communication, a scheme supposedly active in 5m devices today, but is likely too bleeding-edge for our uses. These metadata preservation issues will be further discussed in closing.

OCR vendors increasingly sell powerful document management apps, such as industry leader PaperPort from ScanSoft, whose SDK exports images and such. OCR mainstay Caere's $29.99 PageKeeper is a more open though less used competitor with sophisticated functionality. Other products like Nolo RecordKeeper do much the same for the legal profession, perhaps the biggest user of OCR because many legal searches used to take days. What's in it for Haystack? For mere companions apps, they replicate an amazing amount of Haystack features including typeguessing, querying, and more. They include many powerful format converters that could prove useful to us -- also they offer innovative visual file-browser ideas. Unfortunately however, these products still necessitate "managing your doc manager" and are most useful for organizing visual collections of documents.

3. DESIGN OBJECTIVES

While confining myself to the English language and latin-based characters (though seamless multilingual OCR components often come for free,) this Thesis will attempt to demonstrate that OCR is an empowering personal tool approaching primetime. Whether one uses a desk peripheral or a PDA [see info-capture appliance http://capshare.hp.com] or perhaps a batch process sheet-fed scanner down the hall, a primary goal is to observe individuals and organization moving once-and-for-all across the Rubicon.

Clipped articles from newspapers, popular magazines, and academic journals should, and will, all be scanned into your personal Haystack. All better to help explore how we can better manage our info-clutter portfolios. BusinessWeek [Nov99] claims that Japan will take 6 times as many pictures this year than in 1998, due to the explosion in digital cameras. While personal photos and images are not the center of my project, I intend to build in groundwork for personal [imagery] albums.

Given so much esoterica today comes with a preprinted URL, is it possible that OCR/scanning is now much less necessary? This popular perception must be addressed upfront. It is unlikely because: (1) Most != all. Especially outside the USA, folks are not as URL-crazy and will not be for years. (2) Such self-tagged URLs not always trustworthy: NYT stories often differ offline and online. (3) Magazines like InformationWeek no longer print URLs under each article as they did 3 years ago. (4) Pre-web documents don't have URLs attached. (5) Paper confers many reading-without-surveillance rights that are not preserved online.

Already certain classes of documents are, in fact, designed expressly to defeat OCR "digitization." Certain offline NY Times articles are longer than their free online counterparts. MIT Media Lab members commonly scan artsy business cards onto their web-pages to prevent machine-readability by search engines. Cheap accessible OCR, quite like MP3, would further threaten copyright interests (and privacy), leading to techno-legal restrictions internationally -- even today Xerox is working with the US Secret Service to block scanning of paper currency. In the interim, watermarking and image recognition software are in their infancy, and fostering cultural acceptance digitization should only lead to personal/social workflow empowerment uses. AKA education.

4. CHALLENGES

The nature of this project makes it impossible to anticipate all obstacles. Some, like Paperjam Economics, will be unavoidable. An earnest competitive ergonomics analysis is necessary, even if an early compromise is necessarily made to go forward. A modern digital copier, if well-administered, will avoid much of the incessant paper-fussing of "cheap desktop OCR." Yet users won't OCR (or print) theirs docs at all given paperjams and shortages -- if only for self-cleaning paperpaths!

While more unreliable, users feel more empowered with their own printer/scanner. Much as authoritarian mainframes migrated to the anarchy of personal computation, activators and sensors may do the same. 1996's basic $1000 digital copiers are not only available for $250 -- these multifunction printer/scanners waste far less desk space and you'll never wait for printhogs or leave your cubicle again! Of course, to enterprise sysadmins, supporting ubiquitous "cheap desktop OCR" may present as many headaches as PCs. Bulbs commonly burn out and mechanics fail in today's $39 scanner. Arguably, crumpled paper and staples may not even be the worst of it:

> "The simple task of Verifying the Output of OCR to correct the recognition and formatting errors is always the most cost-intensive component of any OCR application. The raw OCR speed of many hundreds

of pages per hour is limited in the process by the bottleneck of clean-up, which usually proceeds at more human rates than the computer process...the user is presented with a crisp view of each suspect image next to the text in question. With the original and the OCR result viewed side-by-side, the editing process is quick and efficient."

Even vendor Caere here admits to the hidden verification and quality control costs of "spell-checking" and "proofing". Caere has a vested interest --they sell OCR acceleration boards for high volume capture and conversion to stick in the back of your PC. However they are right that expensive network OCR servers used in conjunction with human proofreaders are far more cost-effective, given volume.

How should we maximize ergonomic useability: network scanner and/or desktop scanner and/or handheld? In our case, users should be far more tolerant of errors given "democratic OCR" of personal documents. Still, alternative usage scenarios favor completely different architectures; we might want to support one or two of the following:

(a) Fax-2-email services promise basic OCR. Superb interop if text-only output adequate. Inadequate for small fonts on academic papers.

(b) $99 peripheral + extensible open source? Very immature software :(

(c) $99 peripheral + $99 SW?
Structured HTML output may suffice. Windows GUIs restrictive.

(d) $500 peripheral + $99 SW? Same as (c) with fancier consumer HW:
5ppm combo scanner/page-feeder or snazzy handheld HP Capshare.

(e) $x000 SW? Offers us direct API access to the structure of the document, including mathematical coordinates of words for metadata.
EG. Xerox/Scansoft's TextBridge API v4.5 gives you access to telephone numbers, dates and social security numbers.
At such prices, scanner may "come for free."

(f) $y000 Japanese (or Xerox) digital copier with Ethernet.
Similar to (e) but integrated solution encourages wider adoption.
EG. have HTML (or proprietary pdf) emailed to you before you've even walked back from charmin++ to your office.

Most likely a TWAIN scanner ("Technology Without An Interesting Name") as in (d) will be purchased, though LCS offers a high-quality (non-feeder) flatbed on the 2nd floor. "Buy your software first" is always savvy advice, so especially useful will be newsgroups such as comp.periphs.scanners, alt.comp.periphs.scanner, alt.comp.periphs.scanners and comp.ai.doc-analysis.ocr.

The greatest challenge may be how to exploit (interface with) proprietary software? In our case, we want academic paper abstracts and authors, but how might we incentivize other metadata extraction? And future-proof what we build? OCR'd e-files are semi-structured almost by definition, as recognition algorithms will probably always be one step behind the graphic artists that designed them. EG. $99 COTS (commercial off the shelf) OCR comes with a dozen languages but how do we enable domain-specific plug-ins for esoteric math/science symbology, etc?

Format and filetype matter, unfortunately. The medical industry just spent 8 years arguing over DICOM and SL7 digital image+metadata formats, delaying digitization a decade before finally settling on the same ungainly formats they began with. Mantra: production OCR systems would benefit from open image standards, supplemented with domain-specific XML metadata languages. So what metadata should we use to encapsulate academic papers? A good place to start looking might be IOSIS, the publisher of the Biological Abstracts and Zoological Records, whose production OCR lifts academic papers for a living. Dublin Core politics could be useful too.

One of the main reasons I chose this Thesis topic is that well-defined interfaces on a well-defined problem prevent too-many-cooks. Owning idea(s) and running with them further away from the kitchen (excuse the mixed metaphor) has disadvantages too of course. Working more closely with colleagues on central data model issues would have been both more risky yet more interpersonally fruitful. Which brings me to my final point: that I hope there are still plenty of opportunities for data model design interaction, if only as images are introduced, but elsewhere too.

5. TIMELINE 2000

Jan 15. HW/SW evaluated and structure finalized (block diagram.)
Jan 31. HW/SW delivery begun.

Feb 15. Windows/UNIX fully networked & tested.
Feb 29. Metadata extracted (abstract & authors.)

Mar 15. Metadata integrated into Haystack.
Mar 31. Per-user interfaces polished. Multi-user feedback.

Apr 15. v2.0: users' feedback added.
Apr 30. Thesis done.

May 15. Revised thesis done.
May 31. Future-proofing system, especially HW/SW UIs.

Jun 15. Documentation written.
Jun 30. My birthday!