The Future of Document Imaging in the Era of Electronic Documents

Thomas M. Breuel

Image Understanding and Pattern Recognition Research Group

DFKI and University of Kaiserslautern

www.iupr.org

Electronic Documents Predominate

- Most authoring done with computers
  - MS Office, E-mail, Web, LaTeX, ...
  - text generation: bills, form letters, ...
- Most documents exchanged electronically
  - web (news sites, scientific publications, government publications, public and business forms, ...)
  - email and email attachments
  - groupware and document repositories
- Some laggards
  - books (DRM concerns), legal documents and bill presentment (legal issues, reliability, user reluctance), ...

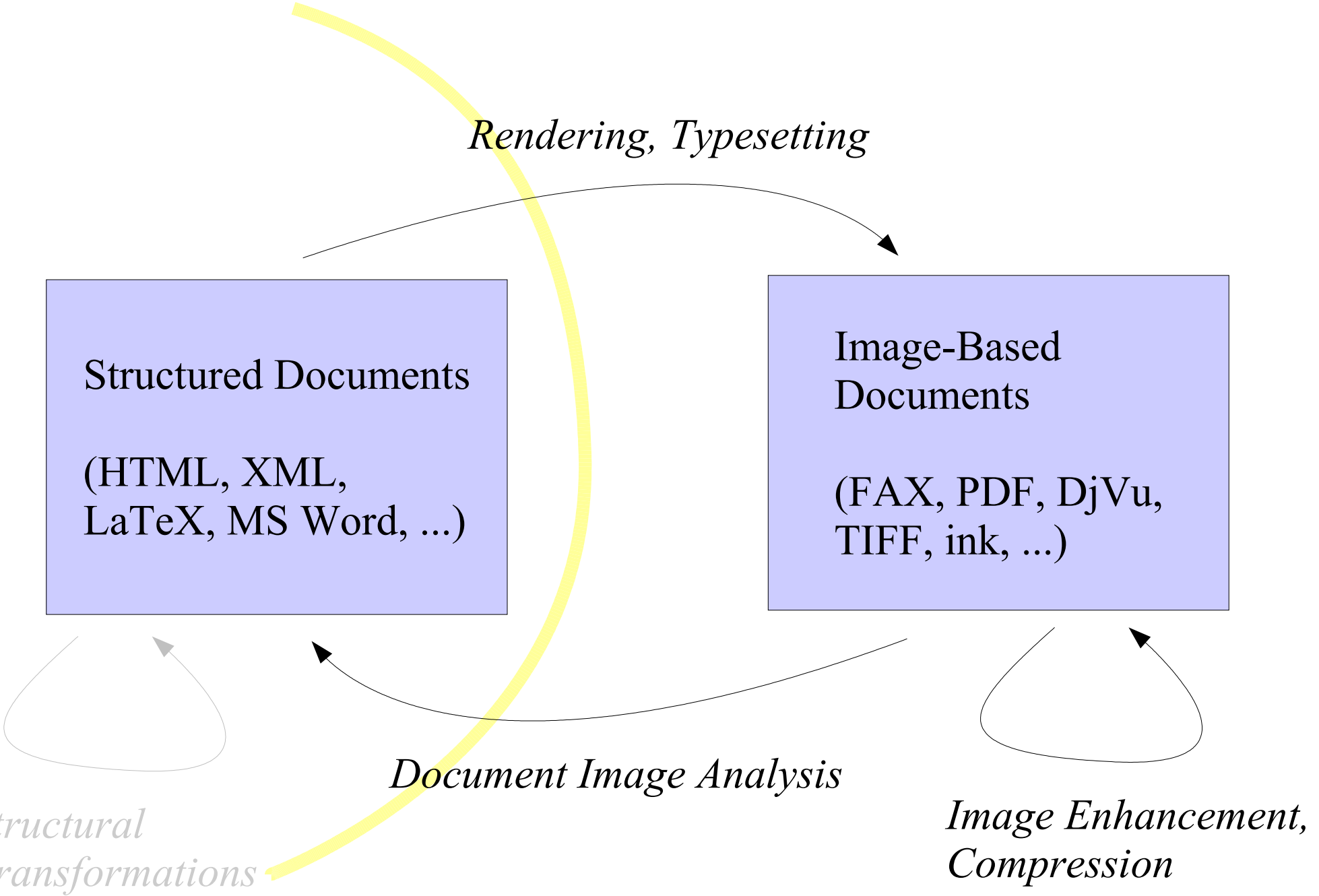
Structured Electronic Documents

- used by office suites, web browsers, presentation packages, forms, ...
- contains
  - the text and its reading order
  - annotations about the logical functions of chunks of text (heading, page number, title, author, etc.)
  - annotations about appearance (italics, bold, font size, etc.)
- semantics of content formally specified
- Examples
  - HTML, XML, LaTeX, MS Word

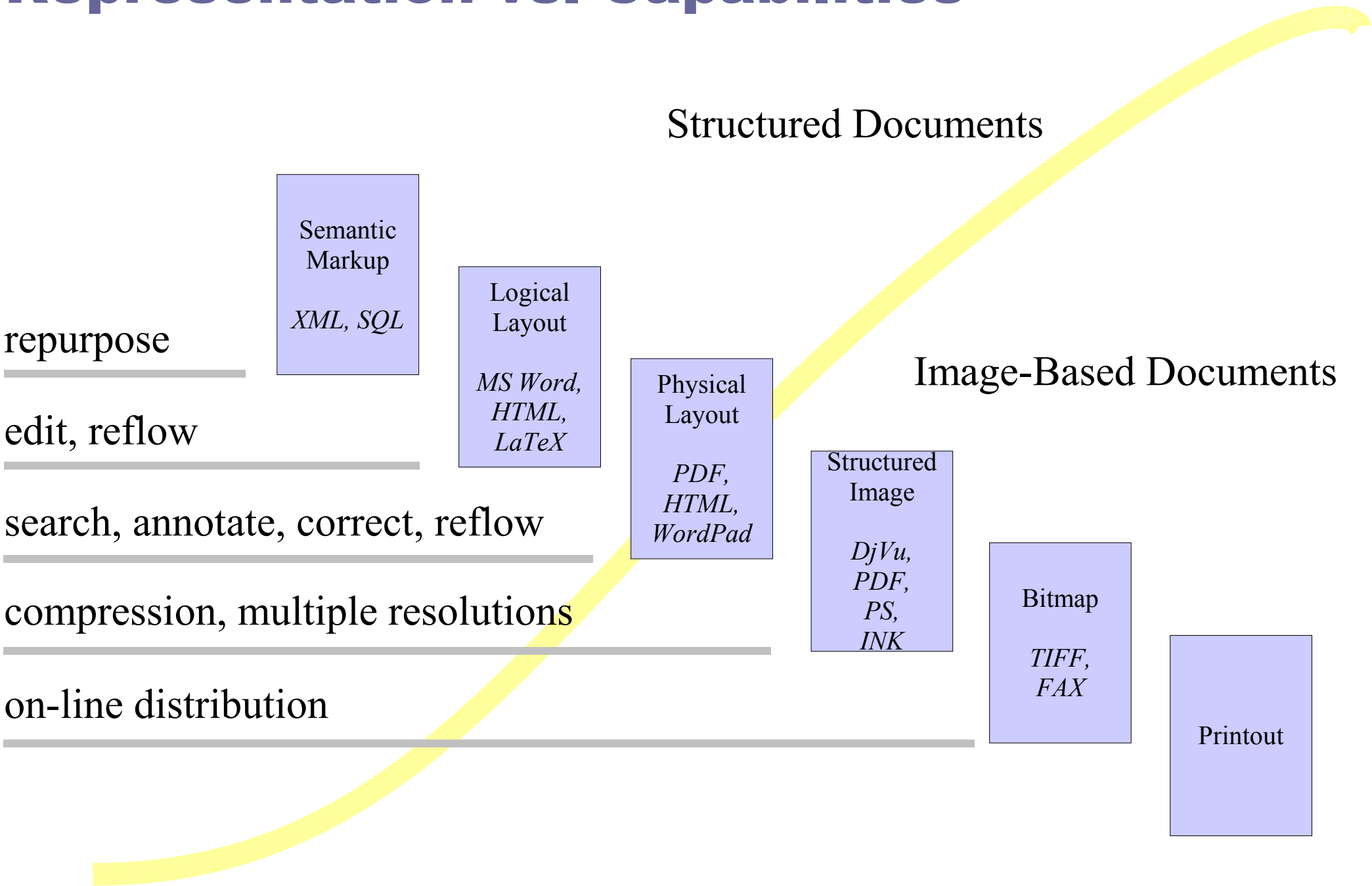
Image-Based Electronic Documents

- obtained by scanning, temporarily created during printing, screen display
- can represent arbitrary images
  - usually pixel-based
    (but could be vector-based, e.g., Ink)
  - little or no information about reading order, logical function
  - may contain text for searching, but the image is what the user sees
- semantics determined by user's interpretation
- Examples
  - TIFF, DjVu, Ink, PDF

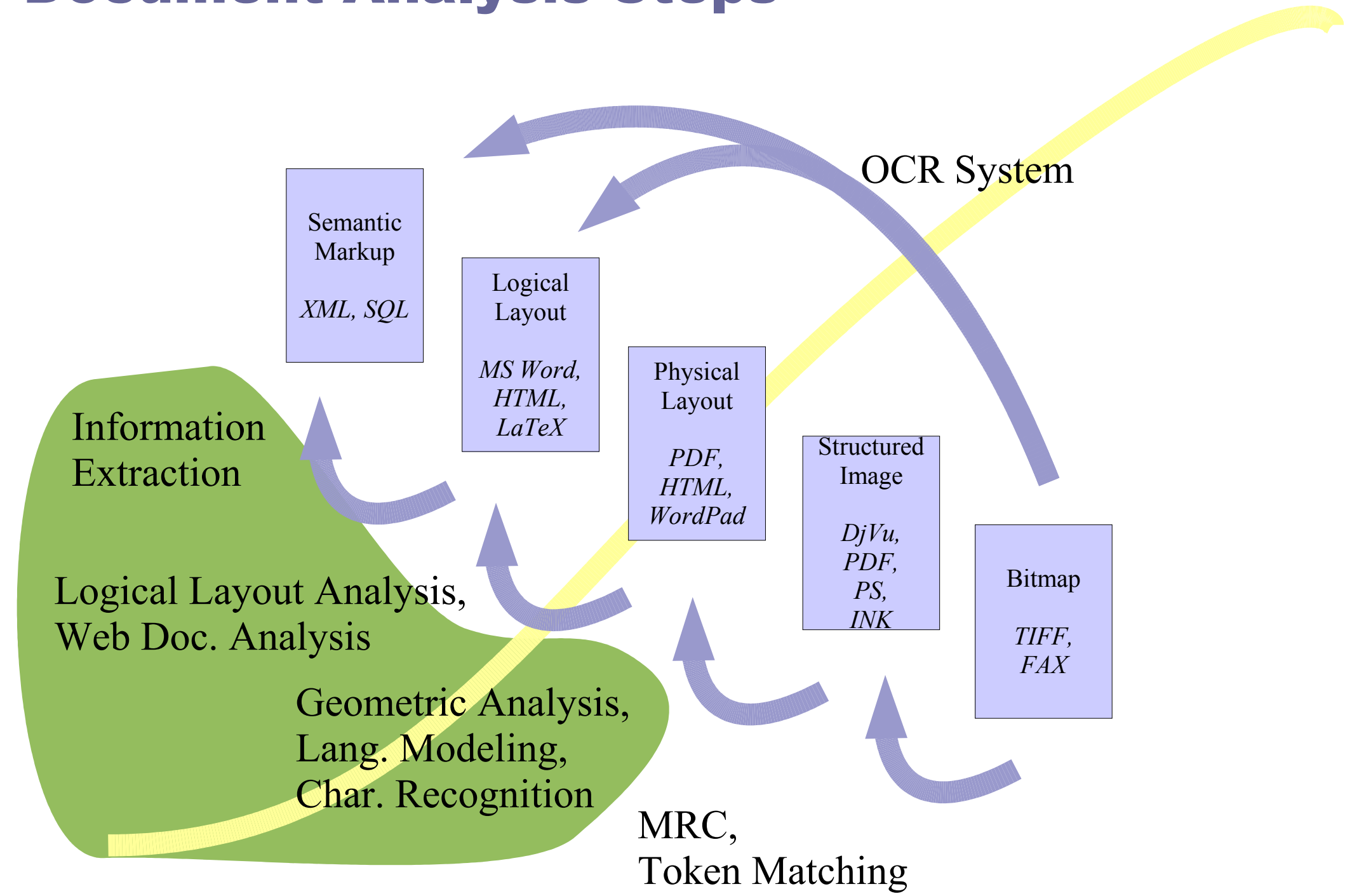
Document Imaging



Representation vs. Capabilities



Document Analysis Steps

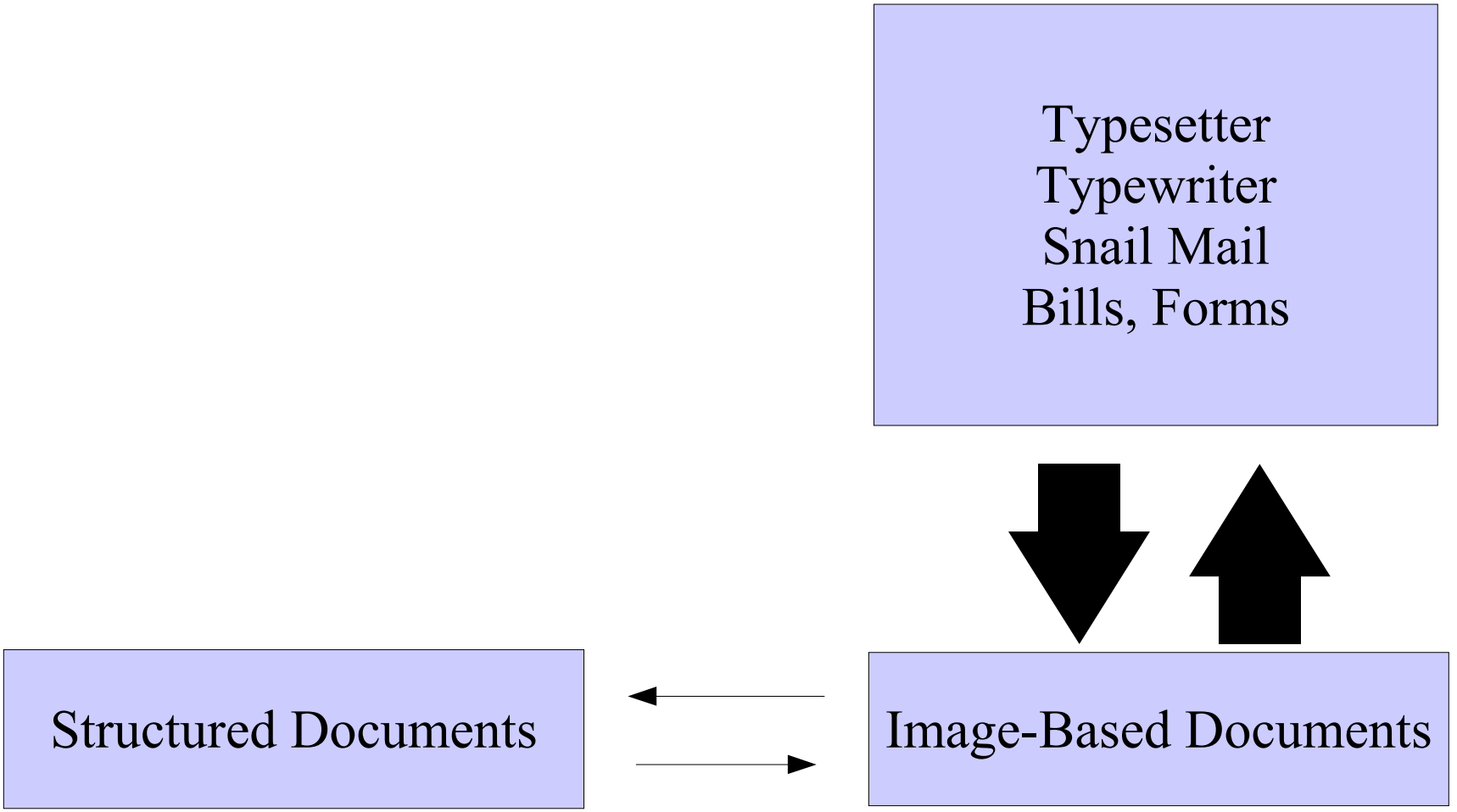


Commonly Made Assumptions

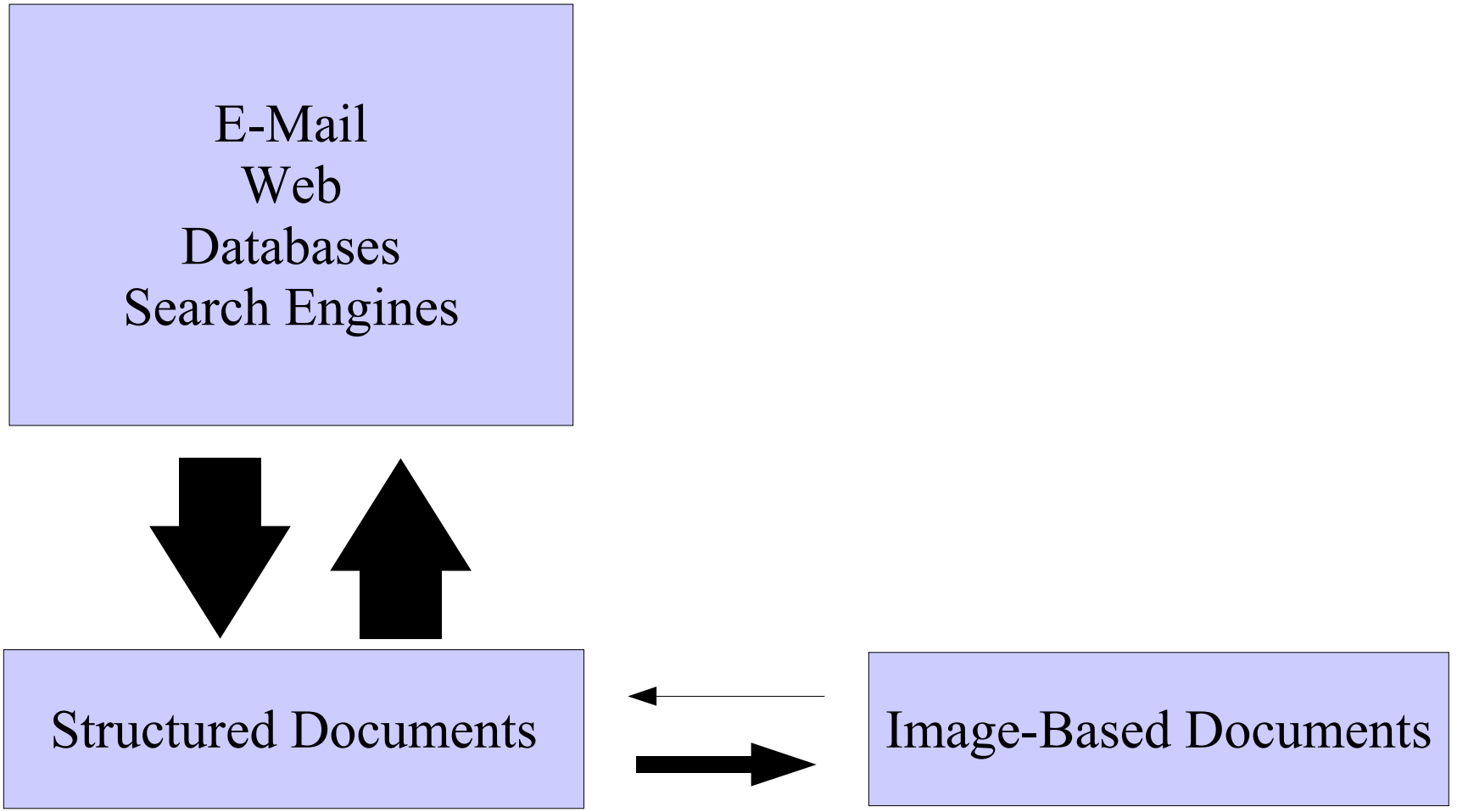
- Paper-based documents are legacy documents—they will gradually be replaced by the web, on-line forms, on-line publishing, electronic handheld readers.
- Semantic, structural, high-level representations of documents are the future.
- Hardcopy documents should be scanned and the resulting Image-based documents should be converted into structured documents as quickly as possible.

- Assumptions drive much of the work behind the web, office suites, OCR, document management, etc.

Traditional Document Use



Assumed Future Document Use



Alternative View

- Paper is not going away for decades to come: a lot of engineering still needs to be done to make electronic displays competitive.
- 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 then should be kept as image-based documents; OCR results should be viewed as an “annotation” of the document image, not as the definitive representation.

Will Paper Go Away?

Will Paper Go Away?

- Arguments eloquently made by Sellen and Harper's book (MIT Press, 2001):
  - **“The Myth of the Paperless Office”**
- “The Paperless Office – In Praise of Clutter” (The Economist, 19 Sep 2002)
- General conclusion:
  - There are some things on-screen reading is really good for, and there are some things paper is really good for
  - Rather than trying to replace one with the other, we should find ways of making them work better together.
  - Presents opportunities for DIA and DAS
- Why? What opportunities?

Hypothetical Electronic Reader Hardware

- Users generally don't like current electronic reader (e-book) hardware and DRM; let's imagine a nearly ideal design...
  - price: $100
  - weight: 0.25kg (0.5 pounds)
  - comm.: IR, Bluetooth, WiFi, UMTS/3G
  - display: 300dpi paper-like display (eInk, Gyricon)
  - format: A6-A4, constrained by user preference
  - storage: 1G flash
  - battery: >24h battery life
  - DRM: no DRM hassles
- (Sony Librie is coming closest, but DRM and price kill it.)
- Probably would be widely sold and widely used for reading.
- The end of paper? The end of document image analysis?

Still Far Away From Paper

- Robustness, Reliability
  - Beach, salt water, bathroom, outdoors cafe, dining hall, airport, babies, dogs, ...
  - Batteries always charged?
- High Cost of Entry, High Risk
  - Lose/damage a $5 book vs. a $100 electronic reader
- Knowledge Work
  - Hard to annotate, can't spread out (need to buy a dozen, but then, how do we manage them?)
- Security/Privacy Issues
  - Do I own the content? Is the content secure? Can I still read my books in 3 years?

Hardware Advances Needed

- Do away with chargers, batteries
  - solar-powered, like some calculators?
- Weight < 50g, Thickness < 5mm for A4/Letter (roll-up an alternative)
  - can carry around several of them
- Cost < $20 / display
  - can afford to buy a lot of them and spread them around
  - loss/damage is less of a risk
- Pen input for annotations
- Fast, robust, nearly invisible operating system
- Sealed, water resistant, flexible

Electronic or Not

- With current hardware...
  - Good Candidates for Electronic Readers
    - big reference works, frequently updated, searchable
    - school textbooks (outdated quickly, heavy)
    - web, email (already electronic)
    - some forms: mail order, government, taxes, etc.
  - Not So Good (print then read instead)
    - location-bound documents/forms: instructions, sign-up information, tourist information, human resource forms, flyers, feedback, handouts, etc.
    - beach reading, travel reading, travel docs
    - technical and scientific papers

Opportunities for DIA in Electronic Readers

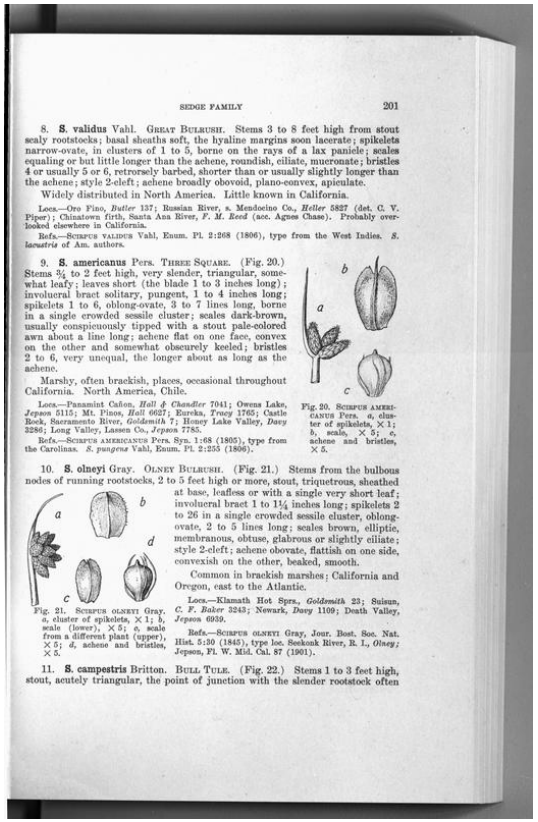
- Strengths/weaknesses point out opportunities for DIA
- Improving Electronic Readers
  - simplify conversion of documents into eBook formats
  - make capture of text into electronic formats easier
  - create long-lived, open document formats
  - transfer some of the capabilities of paper documents over to electronic documents
- Improving Paper/Electronic Integration
  - improve round-trip handling electronic-paper-electronic
    - forms
    - annotations

Improving Electronic Reader Capabilities through DIA

Conversion to eBook Formats ("Paper to PDA", Reflowable Images)

- Idea
  - apply physical layout analysis (detection of text columns, images, reading order, word bounding boxes) to scanned page images
  - use this information to generate a representation of the original document consisting of a mixture of text word images and non-word images in reading order
  - convert the result of the analysis into e-book formats (Plucker, OpenEbook, XHTML, PDF) for display on handheld devices
- Joint work with Baird, Janssen, Popat at PARC

Paper to PDA Examples



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; style 2-cleft; achene broadly obovoid, plano-convex, apiculate.

Widely distributed in North America. Little known in California.

Locs.—Ore Fino, *Butler* 137; Russian River, s. Mendocino Co., *Heller* 5827 (det. C. V. Piper); Chinatown firth, Santa Ana River, *F. M. Reed* (acc. Agnes Chase). Probably overlooked elsewhere in California.

Refs.—SCIRPUS VALIDUS Vahl, Enum. Pl. 2:268 (1806), type from the West Indies. *S. lacustris* of Am. authors.

9. **S. americanus** Pers. THREE SQUARE. (Fig. 20.) Stems ½ to 2 feet high, very slender, triangular, somewhat leafy; leaves short (the blade 1 to 3 inches long); involucral bract solitary, pungent, 1 to 4 inches long; spikelets 1 to 6, oblong-ovate, 3 to 7 lines long, borne in a single crowded sessile cluster; scales dark-brown, usually conspicuously tipped with a stout pale-colored awn about a line long; achene flat on one face, convex on the other and somewhat obscurely keeled; bristles 2 to 6, very unequal, the longer about as long as the achene.

Marshy, often brackish, places, occasional throughout California. North America, Chile.

Locs.—Panamint Cañon, *Hall & Chandler* 7041; Owens Lake, *Jepson* 5115; Mt. Pinos, *Hall* 6627; Eureka, *Tracy* 1765; Castle Rock, Sacramento River, *Goldsmith* 7; Honey Lake Valley, *Davy* 3286; Long Valley, Lassen Co., *Jepson* 7785.

Refs.—SCIRPUS AMERICANUS Pers. Syn. 1:68 (1805), type from the Carolinas. *S. pungens* Vahl, Enum. Pl. 2:255 (1806).

Fig. 20. SCIRPUS AMERICANUS Pers. *a*, cluster of spikelets, × 1; *b*, scale, × 5; *c*, achene and bristles, × 5.

10. **S. olneyi** Gray. OLNEY BULRUSH. (Fig. 21.) Stems from the bulbous nodes of running rootstocks, 2 to 5 feet high or more, stout, triquetrous, sheathed at base, leafless or with a single very short leaf; involucral bract 1 to 1¼ inches long; spikelets 2 to 26 in a single crowded sessile cluster, oblong-ovate, 2 to 5 lines long; scales brown, elliptic, membranous, obtuse, glabrous or slightly ciliate; style 2-cleft; achene obovate, flattish on one side, convexish on the other, beaked, smooth.

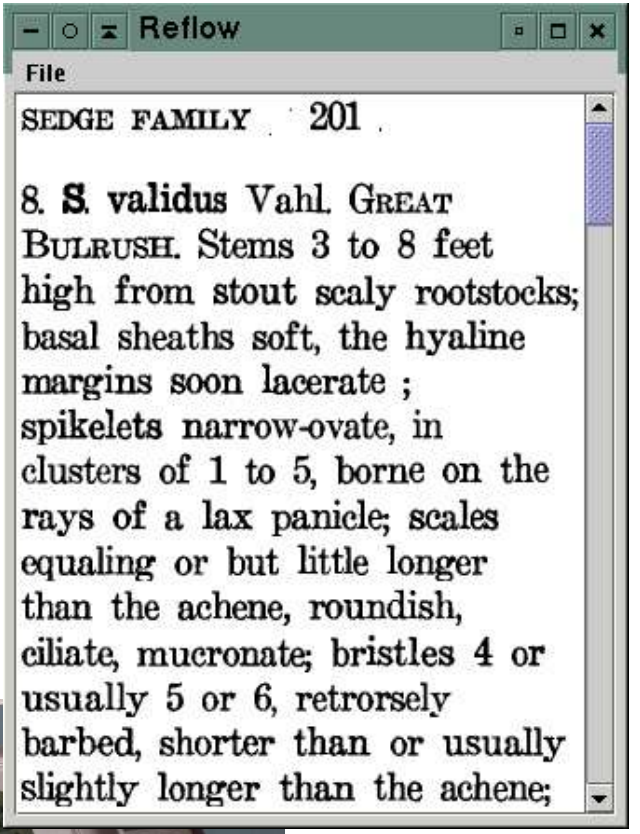
Fig. 21. SCIRPUS OLNEYI Gray. *a*, cluster of spikelets, × 1; *b*, scale (lower), × 5; *c*, scale from a different plant (upper), × 5; *d*, achene and bristles, × 5.

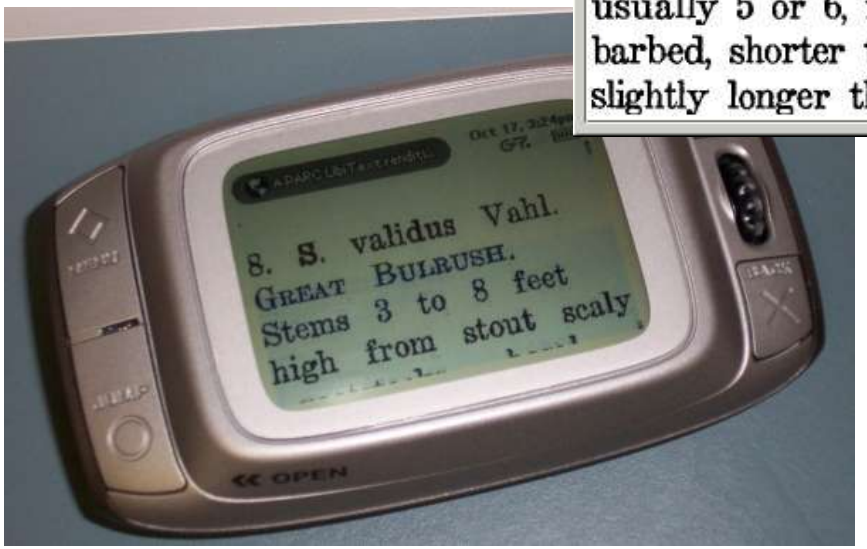
Common in brackish marshes; California and Oregon, east to the Atlantic.

Locs.—Klamath Hot Sprs., *Goldsmith* 23; Suisun, *C. F. Baker* 3243; Newark, *Davy* 1109; Death Valley, *Jepson* 6939.

Refs.—SCIRPUS OLNEYI Gray, Jour. Bost. Soc. Nat. Hist. 5:30 (1845), type loc. Seekonk River, R. I., *Olney*; Jepson, Fl. W. Mid. Cal. 87 (1901).

11. **S. campestris** Britton. BULL TULE. (Fig. 22.) Stems 1 to 3 feet high, stout, acutely triangular, the point of junction with the slender rootstock often





Reflowable Images

- Image-based representation of the original document that can “reflow” when displayed on screens of different sizes.
- Representation is compact and simple: original document image (TIFF, JBIG2, DjVu) plus “reflow annotations”, adding only a few percent to overall size.
- Representation contains the complete, exact original document image
- Representation is universal: can represent any kind of document (math, chemical formulas, ...); no new markup or fonts need to be defined for such content—big advantage over structured e-book formats.
- Compactness, simplicity, preservation of original document image, and universality make it a good candidate for long-lived archival storage.

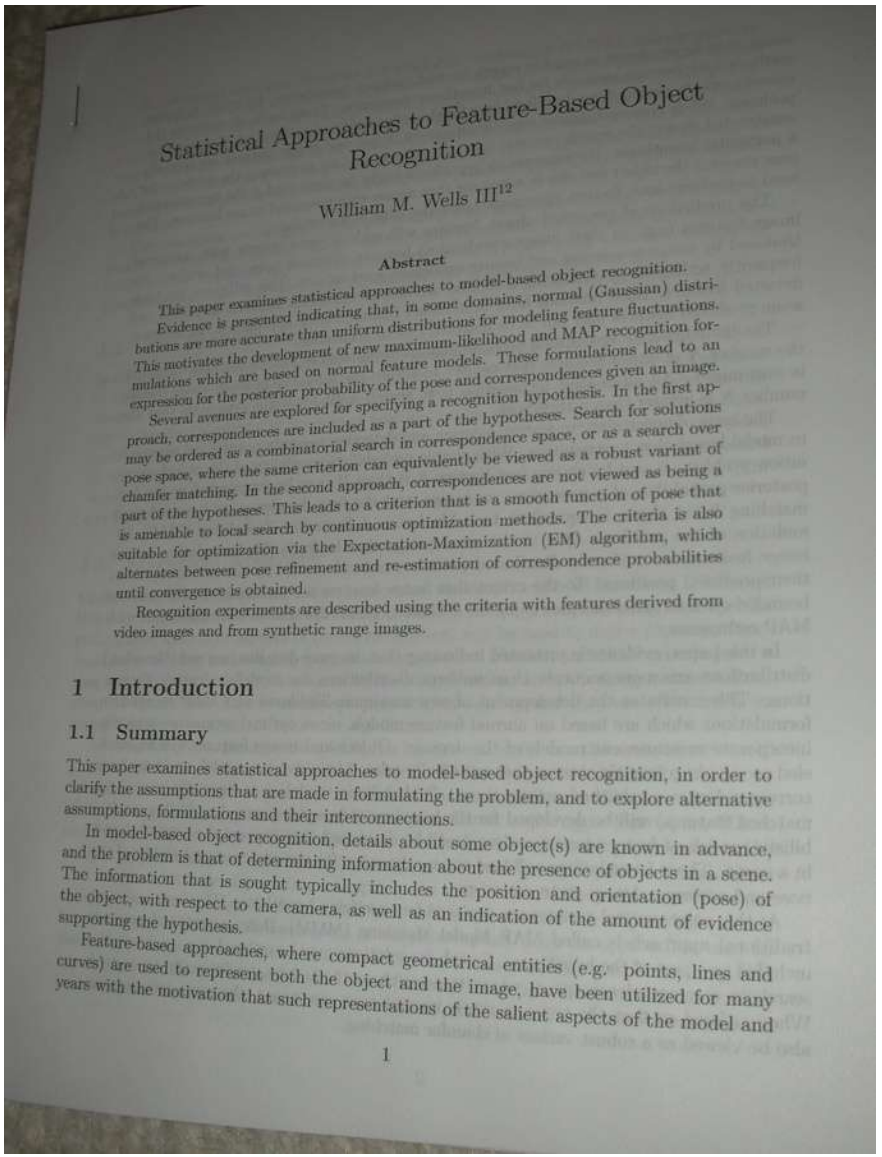
Simplifying Capture

- paper-based documents: can “pick up a flyer” or piece of paper quickly and effortlessly
- attempted alternative: document distribution beacons (e.g., at Sony Metreon, dedicated Palm IR devices)
  - slow, concerns about viruses, compatibility issues
  - not widely deployed, costly to set up/maintain
- camera-based document capture
  - “pick up” a page simply by pointing a camera at it
  - some hands-on experience with Sony DSC-T1: large screen, high resolution, fast operation—yes, it's useful

Camera-Based Document Capture

- recent work in the DIA community
  - Brown and Seales (2001), Cubaud *et al.* (2004), Newman *et al.* (1999), Pilu (2001), Zhang (2004)
- our approaches
  - perspective dewarping using novel text line finder (Breuel, 2002)
    - use reliable text line finder that does not depend on parallelism of text lines
  - stereo-based dewarping (Ulges *et al.,* 2004)
    - use computer vision stereo algorithms, paper shape model to compute 3D document surface
    - use novel dewarping method to dewarp the 3D surface and obtain a flat representation of the document

Perspective Dewarping



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 hypotheses. 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 hypotheses. 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

1

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 hypotheses. 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 hypotheses. 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

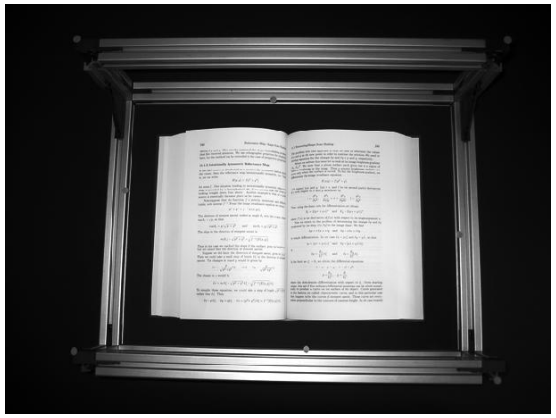
1.1 Summary

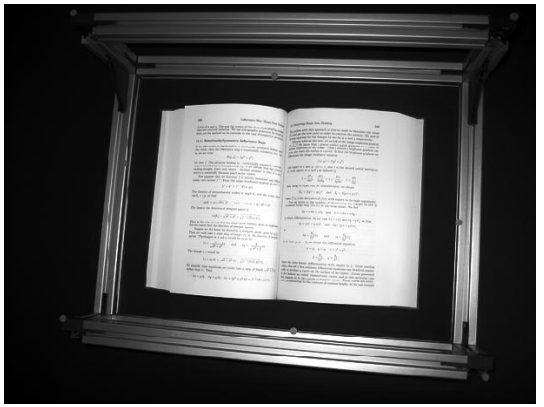
This paper examines statistical approaches to model-based object recognition, in order to clarify the assumptions that are made in formulating the problem, and to explore alternative assumptions, formulations and their interconnections.

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.

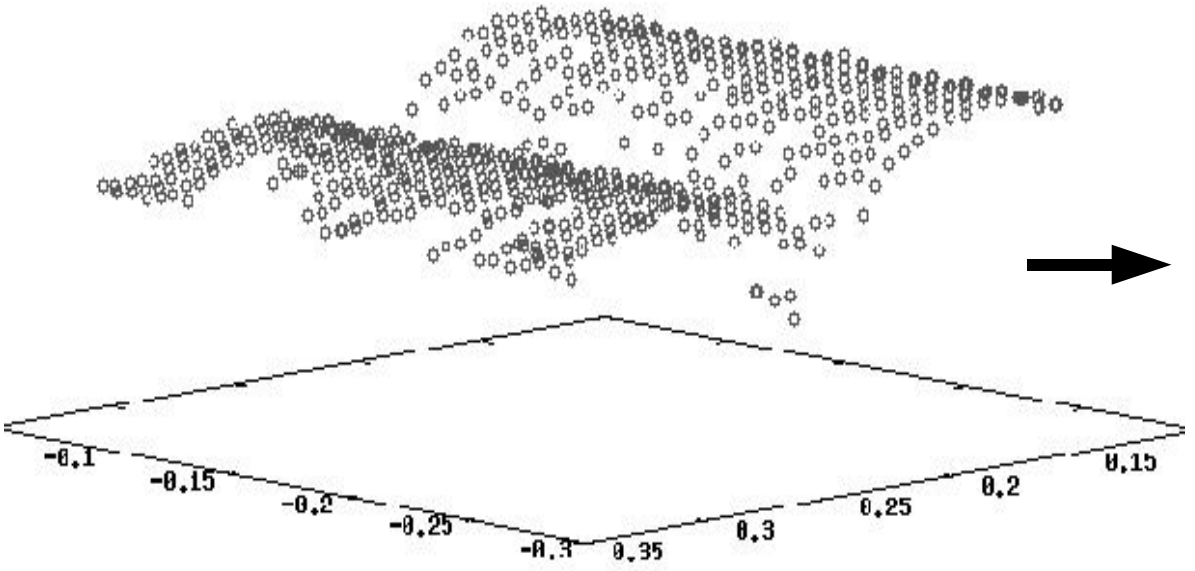
Feature-based approaches, where compact geometrical entities (e.g. points, lines and curves) are used to represent both the object and the image, have been utilized for many years with the motivation that such representations of the salient aspects of the model and

Camera-Based Document Capture w/Stereo





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 a
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
occur only where the surface is curved. To find the
differentiate 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 th
develop equations for the changes δp and δq in p a
Before we address this issue, let us look at the im
$(E_x, E_y)^T$. 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

Camera-Based Document Capture

- allows “grabbing a copy” of a document as easily as pointing the camera at it
- similar to picking up a page, but some differences
  - + physical page can be left behind
  - + fully under recipient's control
  - − time required proportional to number of pages
- stereo capture can be achieved simply by putting two imagers on a device at a few inches apart
- on-going work: improving sparse surface interpolation methods, better dewarping from monocular images

Paper Capabilities to Electronic Documents

- Making documents easier to annotate
  - structured document approach: integrate “electronic ink” data type into documents
    - approach taken by Microsoft Office
  - image-based approaches: add additional layers holding electronic ink information
    - already supported by formats like TIFF
- Making documents easier to “spread out” around the workspace
  - HCI issue, not DIA issue
  - some early work with HCI group, allowing documents to be moved between multiple tablets by spatial gestures

Summary: Improving Electronic Readers

- improve content availability and archival quality of content through developing reflowable image format
- improve ability to acquire/distribute image content for electronic readers through handheld camera-based document capture
- improve ability to annotate content through layered image-based annotations (not shown)
- improve ability to interact with electronic documents by improving the ability to share/distribute/arrange documents among physical devices using spatial gestures
- *Document Imaging is Crucial for making Electronic Reading Devices work in the Real World*

Paper/Electronic Document Integration

Paper/Electronic Document Integration

- “round trip” handling
  - user prints document, works with it, scans it back in, then wants to work with the document electronically again
- “work with” means:
  - archive a high-quality copy of the annotated, scanned page
  - retrieve related electronic versions/documents
  - incorporate marks/annotations/edits on the paper document back into the electronic version
- important special case of “round trip” electronic/paper handling: fill-in forms

Issues Paper/Electronic Document Integration

- Scanning
  - affordable scanners are slow, single-sided, picky about the kind of paper they accept (dog ears etc. lead to misfeeds)
  - without better scanning solutions, users are not going to use round-trip techniques
- Versioning/Robustness
  - existing approaches to paper/electronic document integration fail when the paper and electronic versions of the document differ significantly
- Forms-Related Issues
  - moderate commercial success for fill-in forms, but systems are still too cumbersome for mainstream use

Improved Scanning

- Camera-based capture of documents at the user's desk
  - several previous systems, some commercial products, but quality/performance/success has been limited
- Our approach (on-going work)
  - use new real-time high-resolution cameras to image workspace
  - find less intrusive camera placement, optical path
  - oblivious capture through real-time gesture and activity recognition
  - 3D modeling techniques used for dewarping captured images (as in handheld capture)
- Ultimately: capture everything the user sees
  - leads to wearable computing, but h/w not practical yet

Round-Trip Handling (Annotations)

- several cases
  - separating printed matter from handwritten annotations without knowledge of electronic document
  - detecting annotations assuming the exact electronic version of the document is known
  - comparing electronic document and scanned document even if they are different versions (with layout changes, reflow, etc.)
  - comparing two scanned documents, possibly allowing for layout changes
- some classic document analysis papers cover some of these cases
- we are working on image-based comparisons of different versions of historical documents

Round-Trip Handling (Forms)

- some commercial products that integrate form design tools, scanning, analysis, data capture
- real estate, medical, insurance, etc. prime customers
- based on interviews, users tend not to use these tools
  - form design is often given by external entities (government forms, corporate forms, etc.) and changes frequently
  - conversion of forms into electronic format takes too long, too labor intensive, requires training
  - scanning and image analysis not robust enough
  - users want a mix of on-line and paper capture
- lots of room for improvement in this area through DIA

Long-Term Vision:
Image-Based Personal Computing

Historical Constraints

- focus on structured document types in current computer systems driven by historical constraints
  - costly capture, storage, transmission
  - limited memory, CPU
  - lack of good methods for manipulating/transforming image-based documents
  - bias of computer scientists and software developers towards formal languages, syntactic correctness, precise data models
- structured documents were the only choice users realistically had

Situation Today

- computational infrastructure has changed dramatically
  - image capture, storage, transmission cheap and ubiquitous
  - new sources of image-based documents, e.g. ink from Tablet PC, PDAs
  - memory and CPU fast enough for complex image manipulations, recognition, matching
  - great advances in pattern recognition, document image analysis techniques
  - new generation of computer scientists and software developers looking at adaptive systems, learning, probabilistic systems, pattern recognition
- users can now choose their preferred format
- which one will they choose?

Advantages of Image-Based Docs

Structured Documents

- can represent only content for which markup has been defined
- difficult to convert correctly between different formats
- requires lots of up-front work to create
- by design, contain hidden meta-data—security and privacy issues

Image-Based Documents

- can represent any content (formulas, symbols, ...)
- can be converted among different formats easily and correctly
- once it looks right, you're done
- what you see is what you get, no more, no less
  (unless you deliberately use steganography)

Users Prefer Images/"Analog" in Specific Situations

- Examples
  - use of paper-based mock-up user interfaces in UI design (in preference to dedicated UI prototyping tools)
  - use of handwritten index cards for OO design (in preference to graphical UML design tools)
  - use of Photoshop for web page design, followed by separate manual translation into HTML (in preference to experimenting with designs in HTML directly)
  - use of paper-based forms in preference to on-line forms in many high-tech organizations
- Creative work, brainstorming: preference for tools that impose less structure
- Repetitive work: structure helps

Disadvantages of Image-Based Docs

Structured Documents

- ~~requires little storage~~
- ~~presentation easy to adapt to different screen sizes~~
- presentation can be changed through style files
- allows complex search
- allows data linking/reuse (spreadsheets, addresses, etc.)

Image-Based Documents

- ~~requires lots of storage~~
- ~~presentation difficult to adapt to different screen sizes~~
- presentation difficult to change
- at most, allows keyword search
- no tools for data linking/reuse available (yet)

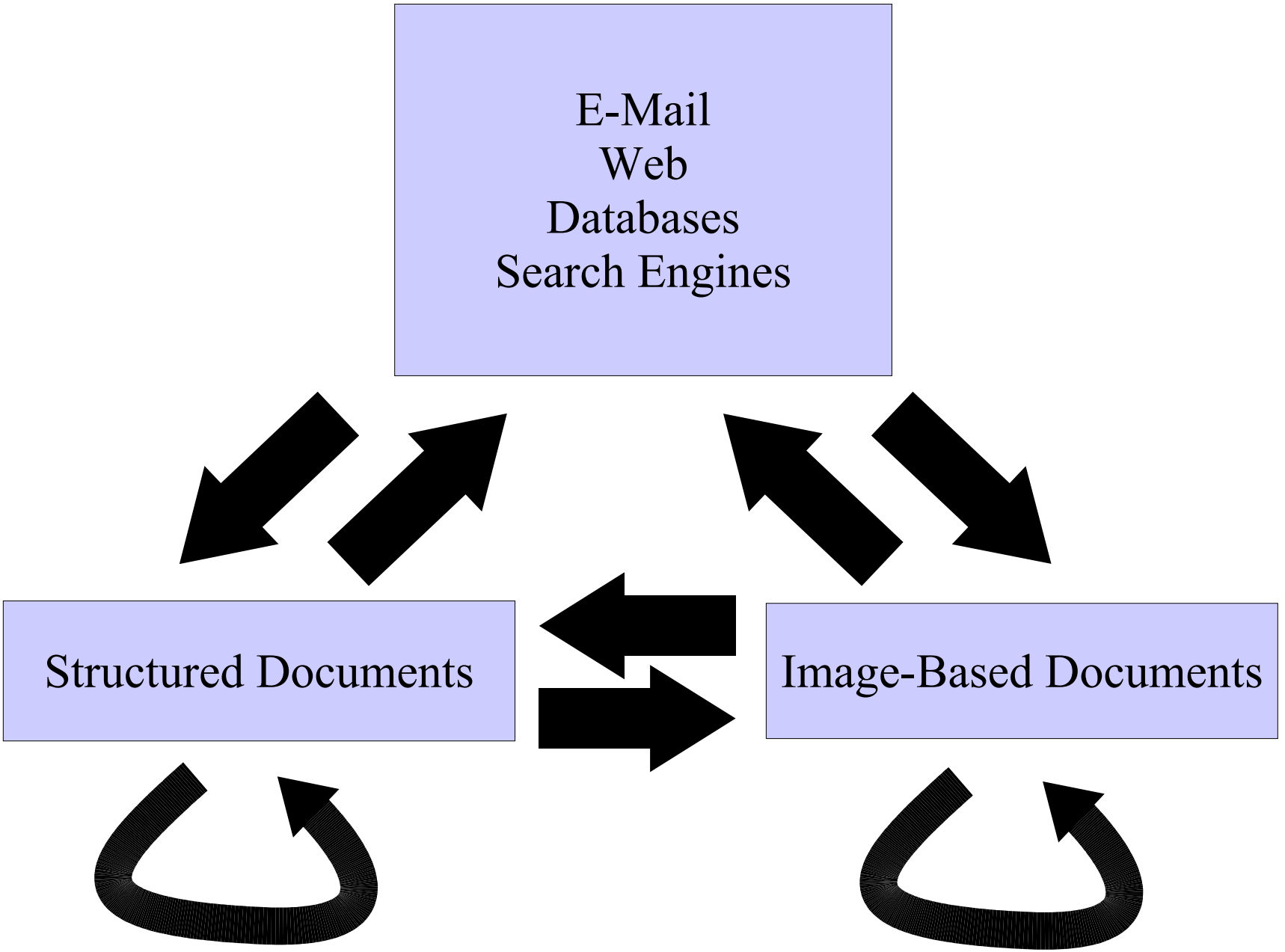
Remaining Disadvantages Fixable?

- e.g.: changing presentation style
  - structured document representation
    - invest time in markup up-front, then change style easily
    - time is wasted if we don't ever change style
  - image-based document representation
    - just get it to look right by whatever means necessary
    - if the style really needs to be changed at some point
      - run it through OCR/layout analysis (easy if clean document)
      - apply style file to output
    - in practice, OCR/layout analysis will probably generate more consistent markup than user anyway
- current practice
  - sufficiently good OCR not widely enough available
  - starting to appear: “auto stylist” in MS Word, StarOffice, etc.
  - other example: Wiki web editing, simple layout analysis of text

Future of Image-Based Representations

- will exist side-by-side with structured representations
- users will pick-and-choose according to task
- operating systems, toolkits need to have easy-to-use support for common conversion/imaging tasks, e.g.:
  - more reliable, robust OCR/layout analysis (commercial offerings not good enough yet: tables, fonts, noise, ...)
  - content-based, appearance-based document image retrieval
  - better abstractions and UI's for image processing tasks
- support for *Image-Based Personal Computing* as extensive as we have it for data types like XML and ASCII

The Future of Documents



Conclusions

- Electronic readers need advances in document imaging technologies in order to succeed even in limited domains.
- Paper is not going away any time soon—it has too many desirable properties—so, traditional document image analysis problems remain.
- Image-based documents and structured documents will increasingly co-exist.
- Document image analysis needs to provide the tools to make this as intuitive and easy for users as possible.