

Towards Reverse Engineering of PDF Documents

Josef Baker, Alan Sexton and Volker Sorge

School of Computer Science, University of Birmingham

July 21, 2011

Motivation

- Accessibility of many scientific PDF documents is poor
 - Poor internal search
 - No integration with other software
- Although many modern articles are published in PDF they rarely (never?) make full use of functionality available in PDF
 - No structure, tags or marked content
- In particular there is no pdf2latex tool!

$$\phi(t) = \frac{1}{\sqrt{2\pi}} \int_0^t e^{-x^2/2} dx$$

□

S 5

,

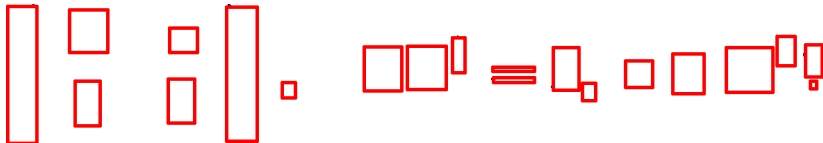
- Overview of previous work
 - Parsing and extraction of formulae from PDF
- Improvements
 - Full document extraction
 - Layout analysis
- Evaluation
 - Comparison to Infty
- Conclusions

- PDF analysis potentially offers more than OCR
- Unicode names, fonts, sizes, baselines are available
- However,

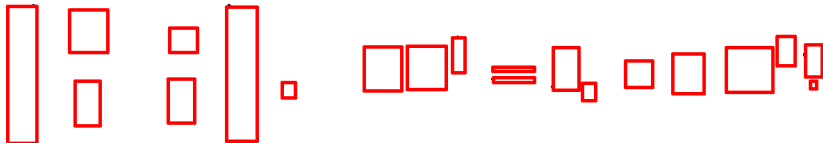
- PDF analysis potentially offers more than OCR
- Unicode names, fonts, sizes, baselines are available
- However,
- Key information may be absent
- Precise spatial information is not available

- PDF analysis potentially offers more than OCR
- Unicode names, fonts, sizes, baselines are available
- However,
- Key information may be absent
- Precise spatial information is not available
- Image analysis also required

- Glyph Extraction



- Glyph Extraction



- PDF Analysis

```
... 10.882 0.199 1 S Q 1 0 0 1 1.307 -9.125 cm BT  
/F11 9.963 Tf 0 0 Td[(k)]TJ/F8 9.963 Tf 5.5 0  
Td[(!)]TJ 5.27 6.834 Td[(050)]TJ/F11 9.963 Tf 3.874 0  
Td[(k)]TJ/F14 9.963 Tf 7.715 0 Td[(000)]TJ/F11 9.963  
Tf 9.962 ...
```


- Linearization

```
matrix(<parenleftbigg, CMEX10, 9.963>)(row(col(<A,  
CMMI10, 9.963>)col(<v, CMMI10, 9.963>))row(col(<zero,  
CMR10, 9.963>)col (<one, CMR10,  
9.963>)))(<parenrightbigg, CMEX10, 9.963>) w3 <comma,  
CMMI10, 9.963> w4 sup <A A, CMMI10, 9.963>(<dagger,  
CMSY7, 6.974>) ...
```

- Linearization

```
matrix(<parenleftbigg, CMEX10, 9.963>)(row(col(<A,  
CMMI10, 9.963>)col(<v, CMMI10, 9.963>))row(col(<zero,  
CMR10, 9.963>)col (<one, CMR10,  
9.963>)))(<parenrightbigg, CMEX10, 9.963>) w3 <comma,  
CMMI10, 9.963> w4 sup <A A, CMMI10, 9.963>(<dagger,  
CMSY7, 6.974>) ...
```

- Parsing and output

$$\begin{pmatrix} A & v \\ 0 & 1 \end{pmatrix}, \quad AA^\dagger = I, \quad v \in \mathbf{R}^3?$$

Improvements: Overview

- Complete page and document extraction
 - No need for manual intervention
 - Suitable for much larger scale conversion
- Structural analysis
 - Math segmentation
 - Layout analysis

Improvements: Extraction

- Extraction and matching extended to whole pages and documents
- Projection Profile Cutting used for line and column detection
 - Efficient and offers good results with many layouts
- Linearization extended for layout analysis
 - Inclusion of line bounding boxes

Improvements: Line Analysis

- Lines are parsed with LALR parser
- Accumulate individual components in each line by
 - assemble single words
 - assemble sequences of mathematical expressions into inline math formulae
- Separate text lines from display style math based on some heuristics (e.g., number of words vs number of math expressions)

Improvements: Assembling Vertical Areas

- Put together paragraphs of parsed lines from previous step plus bounding box information of lines.
- Assemble multiline math expressions by combining consecutive display-style math lines.
- Detect some special features for math paragraphs such as
 - formula enumeration,
 - vertical alignment etc.
- Detect special properties of paragraphs such as
 - alignment, indentation,
 - headers etc.

- Translation into output formats is achieved by specialist drivers
- \LaTeX and MathML drivers for single lines using line analysis information
- \LaTeX driver for entire pages using information on vertical areas plus some spacing information on the layout (MathML still in development).

Original Page

11.7. CONVERGENCE OF SEMIGROUPS.

317

Proof.

$$\begin{aligned}
\|[\exp(n(B-I)) - B^n]x\| &= \left\| e^{-n} \sum_{k=0}^{\infty} \frac{n^k}{k!} (B^k - B^n) x \right\| \\
&\leq e^{-n} \sum_{k=0}^{\infty} \frac{n^k}{k!} \|(B^k - B^n)x\| \\
&\leq e^{-n} \sum_{k=0}^{\infty} \frac{n^k}{k!} \|(B^{k-n} - I)x\| \\
&= e^{-n} \sum_{k=0}^{\infty} \frac{n^k}{k!} \|(B-I)(I+B+\dots+B^{(k-n)-1})x\| \\
&\leq e^{-n} \sum_{k=0}^{\infty} \frac{n^k}{k!} [k-n] \|(B-I)x\|.
\end{aligned}$$

So to prove (11.22) it is enough establish the inequality

$$e^{-n} \sum_{k=0}^{\infty} \frac{n^k}{k!} [k-n] \leq \sqrt{n}. \quad (11.23)$$

Consider the space of all sequences $\mathbf{a} = \{a_0, a_1, \dots\}$ with finite norm relative to scalar product

$$(\mathbf{a}, \mathbf{b}) := e^{-n} \sum_{k=0}^{\infty} \frac{n^k}{k!} a_k \overline{b_k}.$$

The Cauchy-Schwarz inequality applied to \mathbf{a} with $a_k = |k-n|$ and \mathbf{b} with $b_k = 1$ gives

$$e^{-n} \sum_{k=0}^{\infty} \frac{n^k}{k!} [k-n] \leq \sqrt{e^{-n} \sum_{k=0}^{\infty} \frac{n^k}{k!} (k-n)^2} \cdot \sqrt{e^{-n} \sum_{k=0}^{\infty} \frac{n^k}{k!}}.$$

The second square root is one, and we recognize the sum under the first square root as the variance of the Poisson distribution with parameter n , and we know that this variance is n . QED**11.7 Convergence of semigroups.**

We are going to be interested in the following type of result. We would like to know that if A_n is a sequence of operators generating equibounded one parameter semi-groups $\exp tA_n$ and $A_n \rightarrow A$ where A generates an equibounded semi-group $\exp tA$ then the semi-groups converge, i.e. $\exp tA_n \rightarrow \exp tA$. We will prove such a result for the case of contractions. But before we can even formulate the result, we have to deal with the fact that each A_n comes equipped with its own domain of definition, $D(A_n)$. We do not want to make the overly

Rendered L^AT_EX

11.7. CONVERGENCE OF SEMIGROUPS. 317

Proof.

$$\begin{aligned}
\|[\exp(n(B-I)) - B^n]x\| &= \left\| e^{-n} \sum_{k=0}^{\infty} \frac{n^k}{k!} (B^k - B^n) x \right\| \\
&\leq e^{-n} \sum_{k=0}^{\infty} \frac{n^k}{k!} \|(B^k - B^n)x\| \\
&\leq e^{-n} \sum_{k=0}^{\infty} \frac{n^k}{k!} \|(B^{k-n} - I)x\| \\
&= e^{-n} \sum_{k=0}^{\infty} \frac{n^k}{k!} \|(B-I)(I+B+\dots+B^{(k-n)-1})x\| \\
&\leq e^{-n} \sum_{k=0}^{\infty} \frac{n^k}{k!} [k-n] \|(B-I)x\|.
\end{aligned}$$

So to prove (11.22) it is enough establish the inequality

$$e^{-n} \sum_{k=0}^{\infty} \frac{n^k}{k!} [k-n] \leq \sqrt{n}. \quad (11.23)$$

Consider the space of all sequences $\mathbf{a} = \{a_0, a_1, \dots\}$ with finite norm relative to scalar product

$$(\mathbf{a}, \mathbf{b}) := e^{-n} \sum_{k=0}^{\infty} \frac{n^k}{k!} a_k \overline{b_k}.$$

The Cauchy-Schwarz inequality applied to \mathbf{a} with $a_k = |k-n|$ and \mathbf{b} with $b_k = 1$ gives

$$e^{-n} \sum_{k=0}^{\infty} \frac{n^k}{k!} [k-n] \leq \sqrt{e^{-n} \sum_{k=0}^{\infty} \frac{n^k}{k!} (k-n)^2} \cdot \sqrt{e^{-n} \sum_{k=0}^{\infty} \frac{n^k}{k!}}.$$

The second square root is one, and we recognize the sum under the first square root as the variance of the Poisson distribution with parameter n , and we know that this variance is n . QED**11.7 Convergence of semigroups.**

We are going to be interested in the following type of result. We would like to know that if A_n is a sequence of operators generating equibounded one parameter semi-groups $\exp tA_n$ and $A_n \rightarrow A$ where A generates an equibounded semi-group $\exp tA$ then the semi-groups converge, i.e. $\exp tA_n \rightarrow \exp tA$. We will prove such a result for the case of contractions. But before we can even formulate the result, we have to deal with the fact that each A_n comes equipped with its own domain of definition, $D(A_n)$. We do not want to make the overly

- Comparison to Infty's **current** PDF to Latex conversion module
 - Leading scientific mathematical document analysis system
 - Uses commercial OCR software for standard text
 - Specialised OCR for mathematics
 - Performs full page analysis

This is joint work with Masakazu Suzuki [ICDAR 2011]

- 5 scientific papers
 - 2 pages from each
- Wide selection of fonts, maths and layout
- Every page manually ground truthed by Infty
- Required new driver for appropriate output

Evaluation: Character Recognition

- Infty character recognition results

	Artale	Durrett	Judson	Riemann	Sternberg
Objects	11143	3233	1935	2418	2120
Misrecognised	53	5	5	1	3
Extras	46	2	6	2	3
Missing	10	5	4	0	5

- Maxtract character recognition results

	Artale	Durrett	Judson	Riemann	Sternberg
Characters	9304	2799	1744	2094	1889
Symbols	9282	2785	1729	2094	1868
Misrecognised	0	0	0	0	0
Missing	0	0	0	0	0

Evaluation: Formula Recognition

Structure recognition rate wrt. 628 expression.

	Infty	Maxtract
Expression found	635	850
Correct	550	235
Expression split	40	172
Space differences	2	103
Additional characters	10	102
Misrecognised	33	16
Not recognised	7	0

Evaluation: Formula Recognition

Comparison of rendered \LaTeX results

Original	Infty	Maxtract
$r \in \mathbb{N}_{\text{glo}}.$	$r \in \mathbb{N}_{\text{g}^{\text{lo}}}.$	$r \in \mathbb{N}_{\text{ glo }}.$
\mathfrak{J}	$\sim !$	\mathfrak{J}
$\sum_{i=0}^m a_i x^i,$	$\sum_{i=0}^m a_i x^i$	$\sum_{i=0}^m a_i x^i,$

- We have developed a pdf2latex tool
- pdf2mathml also available
- Significant improvements over previous work
 - Now processes entire documents
 - Formulae automatically identified
 - Additional layout analysis
- Layout analysis still naive
- Performs well against leading document analysis system
- Looking forward to results of integration with Infty