A High performance OCR System For Medieval Documents
(e.g. “Narrenschiffe” Novels of 15th Century)

Dr. Saqib Bukhari & Prof. Dr. Andreas Dengel
What we were capable of before the start of this project?
What we were capable of before the start of this project?

We have been developing OCROPUS for a decade

Dr. Syed Saqib Bukhari
What we were cable of before the start of this project?

We have been developing for a decade

Google Sponsored Project to

Complete OCR Work-Flow for Document Analysis

*A project to develop advanced OCR technologies in DFKI*, Google Code Blog, 2007


Preprocessing, Layout Analysis, OCR, ...


Dr. Syed Saqib Bukhari
What we were cable of before the start of this project?

We have been developing [Ocropus](http://ocropus.de) for a decade

*Sponsored Project to Complete OCR Work-Flow for Document Analysis

* "A project to develop advanced OCR technologies in DFKI", Google Code Blog, 2007

Dr. Syed Saqib Bukhari

**Preprocessing, Layout Analysis, OCR, ...**


**OCR Processing for Historical Documents**


What we were capable of before the start of this project?

We have been developing OCRopus for a decade

Google Sponsored Project to Complete OCR Work-Flow for Document Analysis

*A project to develop advanced OCR technologies in DFKI*, Google Code Blog, 2007

Preprocessing, Layout Analysis, OCR, ...

OCR Processing for Historical Documents


So far six PhD theses have been completed on OCRopus, and three are on the way!

Dr. Syed Saqib Bukhari
What we were capable of before the start of this project?

We have been developing OCRopus for a decade

Google Sponsored Project to Complete OCR Work-Flow for Document Analysis

*A project to develop advanced OCR technologies in DFKI*, Google Code Blog, 2007

Preprocessing, Layout Analysis, OCR, ...

OCR Processing for Historical Documents


Ocrosic

Springmann: Ocrosic, a project manager for Ocropus, Ludwig-Maximilians-University, Munich, 2015.

So far six PhD theses have been completed on OCRopus, and three are on the way!

Dr. Syed Saqib Bukhari
What we were capable of before the start of this project?

An Example: The Work-Flow For 15th Century Historical Documents

Adam von Bodenstein (1557) [1]

What we were capable of before the start of this project?

An Example: The Work-Flow For 15th Century Historical Documents

Adam von Bodenstein (1557) [1]

Preprocessing and Layout Analysis

What we were capable of before the start of this project?

An Example: The Work-Flow For 15th Century Historical Documents

Adam von Bodenstein (1557) [1]

Preprocessing and Layout Analysis

Trained OCR Line Recognizer

OCRopus Results [1]

Dr. Syed Saqib Bukhari

What we were capable of before the start of this project?

An Example: The Work-Flow For 15th Century Historical Documents

On Test images from Adam von Bodenstein (1557), the performance:

**Ocropus 99%**, **ABBYY 85%** and and **Tesseract 78%** [1].

What we were capable of before the start of this project?

An Example: The Work-Flow For 15th Century Historical Documents

On Test images from Adam von Bodenstein (1557), the performance: Ocropus 99%, ABBYY 85% and and Tesseract 78% [1].

so what else is required for the 15th Century “Narrenschiffe” novels of this project?
What are the limitations of the OCRopus?
What are the limitations of the OCRopus?

Layout Analysis
What are the limitations of the OCRopus?

Layout Analysis

OCRopus 100%
What are the limitations of the OCRopus?

Layout Analysis

OCRopus 100%

OCRopus 75%

Dr. Syed Saqib Bukhari
What are the limitations of the OCRopus?

Layout Analysis

OCRopus 100%

XY Cut < 5%

Voronoi < 5%

OCRopus 75%

XY Cut < 5%

Voronoi < 5%

OCRopus 25%

XY Cut < 5%

Voronoi < 5%

*OCRBukhari*
What are the limitations of the OCRopus?

Layout Analysis

What are we dealing in this project?
What are the limitations of the OCRopus?

Layout Analysis

What are we dealing in this project?

so what else is required for the 15th Century “Narrenschiffe” novels of this project?

(i) Advanced Layout Analysis
What are the limitations of the OCRopus?

OCR Model
What are the limitations of the OCRopus?

OCR Model

[Diagram showing a flow of data from pixel values to recognized labels and a text-line image with associated ground-truth file.]

Dr. Syed Saqib Bukhari
What are the limitations of the OCRopus?

OCR Model

Adam von Bodenstein 1557 [1]
(OCRopus: 99%, ABBYY: 85%. Tesseract: 78% )

Augustinus Leyer 1735 [1]
(OCRopus: 97%, ABBYY: 77%. Tesseract: 82% )

Augsburg before 1476 [1]
(OCRopus: 98%)


Dr. Syed Saqib Bukhari
What are the limitations of the OCRopus?

OCR Model

Training Data
50K Lines

Adam von Bodenstein 1557 [1]
(OCRopus: 99%, ABBYY: 85%. Tesseract: 78%)

Augustinus Leyer 1735 [1]
(OCRopus: 97%, ABBYY: 77%. Tesseract: 82%)

Augsburg before 1476 [1]
(OCRopus: 98%)


Dr. Syed Saqib Bukhari
What are the limitations of the OCRopus?

**OCR Model**

<table>
<thead>
<tr>
<th>German</th>
<th>Latin</th>
<th>French</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image1.jpg" alt="Image" /></td>
<td><img src="image2.jpg" alt="Image" /></td>
<td><img src="image3.jpg" alt="Image" /></td>
</tr>
<tr>
<td><img src="image4.jpg" alt="Image" /></td>
<td><img src="image5.jpg" alt="Image" /></td>
<td><img src="image6.jpg" alt="Image" /></td>
</tr>
<tr>
<td><img src="image7.jpg" alt="Image" /></td>
<td><img src="image8.jpg" alt="Image" /></td>
<td><img src="image9.jpg" alt="Image" /></td>
</tr>
</tbody>
</table>

What are we dealing in this project?
What are the limitations of the OCRopus?

**OCR Model**

What are we dealing in this project?

<table>
<thead>
<tr>
<th>German</th>
<th>Latin</th>
<th>French</th>
</tr>
</thead>
<tbody>
<tr>
<td>Basel 1495 (5046)</td>
<td>Basel 1497 (5054)</td>
<td>Lyon 1499 (5060)</td>
</tr>
<tr>
<td>Basel 1499 (5047)</td>
<td>Basel 1498 (5062)</td>
<td>Paris 1499 (5065)</td>
</tr>
</tbody>
</table>

Dr. Syed Saqib Bukhari
What are the limitations of the OCRopus?

OCR Model

What are we dealing in this project?

<table>
<thead>
<tr>
<th>German</th>
<th>Latin</th>
<th>French</th>
</tr>
</thead>
<tbody>
<tr>
<td>Basel 1495 (5046)</td>
<td>Basel 1497 (5054)</td>
<td>Lyon 1499 (5060)</td>
</tr>
<tr>
<td>Basel 1499 (5047)</td>
<td>Basel 1498 (5062)</td>
<td>Paris 1499 (5065)</td>
</tr>
</tbody>
</table>

so what else is required for the 15th Century “Narrenschiffe” novels of this project?

(ii) Automatic OCR Model (anyOCR!)
Our Main Goals in This Project!
Our Main Goals in This Project!

**Advanced Layout Analysis**

**Automatic OCR Model (anyOCR)**
Our Main Goals in This Project!

**Advanced Layout Analysis**

**Automatic OCR Model (anyOCR)**
Our Main Goals in This Project!

Advanced Layout Analysis

Automatic OCR Model (anyOCR)
Advanced Layout Analysis

Text and Non-Text Segmentation Method
Advanced Layout Analysis

Text and Non-Text Segmentation Method

Dr. Syed Saqib Bukhari
Advanced Layout Analysis

Text Line Segmentation

XY Cut

Voronoi

OCRopus
Advanced Layout Analysis

Text Line Segmentation

XY Cut  Voronoii  OCRopus  OCRopus++
Performance Evaluation of **OCRopus++**: 

- **Text and Non-Text Segmentation Accuracy**: 99.34%
- **Text Line Segmentation Accuracy**: 87%* as compared to state-of-the-art methods ~10% and Ocropus 80%.
Performance Evaluation of **OCRopus++**: 

- **Text and Non-Text Segmentation Accuracy:** 99.34%
- **Text Line Segmentation Accuracy:** 87%*  
*as compared to state-of-the-art methods ~10% and OCRopus 80%.

Dr. Syed Saqib Bukhari
Performance Evaluation of **OCRopus++**: 

- **Text and Non-Text Segmentation Accuracy**: 99.34%
- **Text Line Segmentation Accuracy**: 87%*  

*as compared to state-of-the-art methods ~10% and **OCRopus** 80%.
Performance Evaluation of **OCRopus++**:  

- **Text and Non-Text Segmentation Accuracy**: 99.34%  
- **Text Line Segmentation Accuracy**: 87%*  

*as compared to state-of-the-art methods ~10% and OCRopus 80%.
Our Main Goals in This Project!

Advanced Layout Analysis

Automatic OCR Model (anyOCR)

Dr. Syed Saqib Bukhari
Background: OCR Training Models can broadly be classified as:

- **Segmentation-based OCR (Tesseract)**
  - individual characters classification

- **Segmentation-Free OCR (OCRopus)**
  - line recognizer
Background: OCR Training Models can broadly be classified as:

- **Segmentation-based OCR (Tesseract)**
  - individual characters classification

- **Segmentation-Free OCR (OCRopus)**
  - line recognizer

Tools: Aletheia and Franken+

Training Data

~50k to 100k
OCR Training Models can broadly be classified as:

- **Segmentation-based OCR (Tesseract)**
  - individual characters classification

- **Segmentation-Free OCR (OCRopus)**
  - line recognizer

Tools:
- Aletheia
- Franken+

Training Data:
- ~50k to 100k

---

Dr. Syed Saqib Bukhari
**Background:** OCR Training Models can broadly be classified as:

- **Segmentation-based OCR** *(Tesseract)*
  - individual characters classification
  - *Lower Performance*
    - Sensitive to Noise
    - Character Segmentation Errors
    - No Language Model

- **Segmentation-Free OCR** *(OCRopus)*
  - line recognizer
  - *Higher Performance*
    - Insensitive to Noise
    - Independent to Noise & Character Seg.
    - Implicit Language Model

---

**Tools:** Aletheia and Franken+

**Training Data:**

<table>
<thead>
<tr>
<th>TypeTable / Few Pages</th>
</tr>
</thead>
<tbody>
<tr>
<td>~50k to 100k</td>
</tr>
</tbody>
</table>

Dr. Syed Saqib Bukhari
Background: OCR Training Models can broadly be classified as:

- **Segmentation-based OCR (Tesseract)**
  - individual characters classification
  - **Lower Performance**
    - Sensitive to Noise
    - Character Segmentation Errors
    - No Language Model

- **Segmentation-Free OCR (OCRopus)**
  - line recognizer
  - **Higher Performance**
    - Insensitive to Noise
    - Independent to Noise & Character Seg.
    - Implicit Language Model

Training Data:
- TypeTable / Few Pages
  - Training Data: Roma fenes habuit consultoreg peritos
    - Roma fenes habuit consultoreg peritos
  - Nil libi diuitie profunt: nil fortis auare
    - Nil libi diuitie profunt: nil fortis auare
  - Optaut nemo lapiens/ qui in orbe potenter
    - Optaut nemo lapiens/ qui in orbe potenter

Tools: Aletheia and Franken+

Dr. Syed Saqib Bukhari
anyOCR - Automatic OCR Model

Background: OCR Training Models can broadly be classified as:

- **Segmentation-based OCR (Tesseract)**
  - individual characters classification
  - Lower Performance
  - Sensitive to Noise
  - Character Segmentation Errors
  - No Language Model

- **Segmentation-Free OCR (OCRopus)**
  - line recognizer
  - Higher Performance
  - Insensitive to Noise
  - Independent to Noise & Character Seg.
  - Implicit Language Model

Tools:
- Aletheia and Franken+
- TypeTable

Training Data:
- Adam von Bodenstein 1557
  - Ocropolis: 99%, Tesseract: 78%

Dr. Syed Saqib Bukhari
anyOCR: A Combination of OCRopus and tesseRECT (OCRoRECT)

- **Training Data:** only TypeTable

- **Higher Performance:**
  - Comparable to OCRopus
  - Independent to Noise & Character Seg.
  - Implicit Language Model
**anyOCR** - Automatic OCR Model

**anyOCR**: A *Combination of OCRopus and tesseRECT* (**OCRoRECT**)

- **Training Data**: only TypeTable

- **Higher Performance**:
  - Comparable to OCRopus
  - Independent to Noise & Character Seg.
  - Implicit Language Model
anyOCR: A Combination of OCRopus and tesseRECT (OCRoRECT)

- **Training Data**: only TypeTable

- **Higher Performance**:
  - Comparable to OCRopus
  - Independent to Noise & Character Seg.
  - Implicit Language Model
anyOCR: A Combination of OCRopus and tesseRECT (OCRoRECT)

- **Training Data:** only TypeTable

- **Higher Performance:**
  - Comparable to OCRopus
  - Independent to Noise & Character Seg.
  - Implicit Language Model
anyOCR: A Combination of OCRopus and tesseRECT (OCRoRECT)

- **Training Data:** only TypeTable

- **Higher Performance:**
  - Comparable to OCRopus
  - Independent to Noise & Character Seg.
  - Implicit Language Model

---

Dr. Syed Saqib Bukhari
anyOCR - Automatic OCR Model

**anyOCR**: A Combination of **OCRopus** and **tesserECT** (**OCRoRECT**)

- **Training Data**: only TypeTable
- **Higher Performance**:
  - Comparable to OCRopus
  - Independent to Noise & Character Seg.
  - Implicit Language Model
anyOCR: A Combination of OCRopus and tesseRECT (OCRoRECT)

- **Training Data:** only TypeTable

- **Higher Performance:**
  - Comparable to OCRopus
  - Independent to Noise & Character Seg.
  - Implicit Language Model

Dr. Syed Saqib Bukhari
anyOCR: A Combination of OCRopus and tesseRECT (OCRoRECT)

- **Training Data:** only TypeTable
- **Higher Performance:**
  - Comparable to OCRopus
  - Independent to Noise & Character Seg.
  - Implicit Language Model

---

Dr. Syed Saqib Bukhari
anyOCR: A Combination of OCRopus and tesseRECT (OCRoRECT)

- **Training Data:** only TypeTable

- **Higher Performance:**
  - Comparable to OCRopus
  - Independent to Noise & Character Seg.
  - Implicit Language Model

---

**anyOCR:**

- **Training Data:** only TypeTable
- **Higher Performance:**
  - Comparable to OCRopus
  - Independent to Noise & Character Seg.
  - Implicit Language Model

---


---

Dr. Syed Saqib Bukhari
anyOCR: A Combination of OCRopus and tessereCT (OCRoRECT)

- **Training Data**: only TypeTable
- **Higher Performance**:
  - Comparable to OCRopus
  - Independent to Noise & Character Seg.
  - Implicit Language Model

anyOCR: A Combination of OCRopus and tesseRECT (OCRoRECT)

- **Training Data:** only TypeTable
- **Higher Performance:**
  - Comparable to OCRopus
  - Independent to Noise & Character Seg.
  - Implicit Language Model

---

anyOCR - Automatic OCR Model

anyOCR: A Combination of OCRopus and tesseract (OCRoRECT)

- **Training Data:** only TypeTable
- **Higher Performance:**
  - Comparable to OCRopus
  - Independent to Noise & Character Seg.
  - Implicit Language Model

---

Dr. Syed Saqib Bukhari

---

Meaningless Text Generator

Lufticiq; modum: quo/re sapienter in omni
Consilium monstrare tuum: mentemq3 valeres:
Antea consuleres doctos/iurifq3 peritos:

Correct Text Lines

Lufticiq; modum: quo/re sapienter in omni
Consilium monstrare tuum: mentemq3 valeres:
Antea consuleres doctos/iurifq3 peritos:

Erroneous Text Lines

Line Images

Ocropus Training

Tesseract Training

[Deutsches Forschungszentrum für Künstliche Intelligenz GmbH]
Qualitative Performance Evaluation of OCRopus++

\[\text{Iusticiæ modum: quo\textit{re} sapienter in omni} \]
\[\text{Consilium monstrare tuum: mentemq3 valeres:} \]
\[\text{Antea consulures do\textit{ctos/iurifq3 peritos:}} \]

Tesseract

anyOCR: OCRoRECT

Dr. Syed Saqib Bukhari
Qualitative Performance Evaluation of **OCRopus++**

Iūticiēsmodum:quo/re sapienter in omni
Consilium monstrare tuum:mentemq3 valeres:
Antea consuleres doctos/iurifq3 peritos:

Tesseract  OCRopus  anyOCR: OCRoRECT
Performance Evaluation of **OCRopus++**:

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Tesseract</th>
<th>OCRopus</th>
<th>anyOCR - OCRoRECT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adam von Bodenstein 1557</td>
<td>78% [1]</td>
<td>99% [1]</td>
<td>-</td>
</tr>
<tr>
<td>Augustinus Leyer 1735</td>
<td>82% [1]</td>
<td>97% [1]</td>
<td>-</td>
</tr>
</tbody>
</table>

Performance Evaluation of **OCRopus++**:  

<table>
<thead>
<tr>
<th>Dataset</th>
<th>OCR Model</th>
<th>Tesseract</th>
<th>OCRopus</th>
<th>anyOCR - OCRoRECT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adam von Bodenstein 1557</td>
<td></td>
<td>78% [1]</td>
<td>99% [1]</td>
<td>-</td>
</tr>
<tr>
<td>Augustinus Leyer 1735</td>
<td></td>
<td>82% [1]</td>
<td>97% [1]</td>
<td>-</td>
</tr>
<tr>
<td>Basel 1497 (Narrenschif, Latin)</td>
<td></td>
<td>77% [2]</td>
<td>98% [2]</td>
<td>-</td>
</tr>
</tbody>
</table>

### Performance Evaluation of OCRopus++:

<table>
<thead>
<tr>
<th>Dataset</th>
<th>OCR Model</th>
<th>Tesseract</th>
<th>OCRopus</th>
<th>anyOCR - OCRoRECT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adam von Bodenstein 1557</td>
<td>Tesseract</td>
<td>78% [1]</td>
<td>99% [1]</td>
<td>-</td>
</tr>
<tr>
<td>Augustinus Leyer 1735</td>
<td>Tesseract</td>
<td>82% [1]</td>
<td>97% [1]</td>
<td>-</td>
</tr>
</tbody>
</table>

Conclusion
Conclusion

- Layout Analysis is an open-challenging problem for complex documents.

- OCR recognizers those need less training data fail to achieve good performance (e.g. Tesseract) and those produce better results require a lot of training data (e.g. OCRopus).
Conclusion

- Layout Analysis is an open-challenging problem for complex documents

- OCR recognizers those need less training data fail to achieve good performance (e.g. Tesseract) and those produce better results require a lot of training data (e.g. OCRopus)

- Presented OCRopus++ OCR System
  - advanced layout analysis
  - anyOCR - OCRoRECT recognizer
Conclusion

- Layout Analysis is an open-challenging problem for complex documents

- OCR recognizers those need less training data fail to achieve good performance (e.g. Tesseract) and those produce better results require a lot of training data (e.g. OCRopus)

- Presented **OCRopus++** OCR System
  - advanced layout analysis
  - anyOCR - OCRoRECT recognizer

- In future, add more features in **OCRopus++**
Conclusion

● Layout Analysis is an open-challenging problem for complex documents

● OCR recognizers those need less training data fail to achieve good performance (e.g. Tesseract) and those produce better results require a lot of training data (e.g. OCRopus)

● Presented OCRopus++ OCR System
  ○ advanced layout analysis
  ○ anyOCR - OCRoRECT recognizer

● In future, add more features in OCRopus++

● Publications:
Thank you, ... Questions?

Address:
Dr.-Ing. Syed Saqib Bukhari
DFKI GmbH
P.O. Box 2080
D–67608 Kaiserslautern
e-mail: saqib.bukhari@dfki.de
http://www.dfki.de/~bukhari