PDF Days Europe 2025

Logical structure rediscovery for automatic PDF tagging

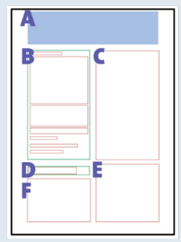
Rule-based, machine learning and generative AI approaches

Tamir Hassan • Co-Founder and CTO, Living PDF • tamir@tamirhassan.com

Living PDF

Pipeline

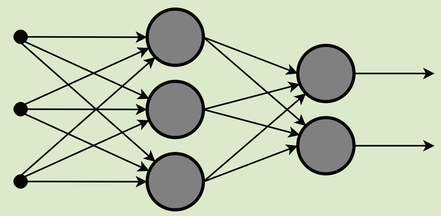
Includes OCR (if required), segmentation, logical labelling, article and reading order detection and alt text generation




```
<?xml version="1.0" encoding="UTF-8"?>
<document>
  <textual-content id="1" style-id="1">
    <p><b>Liebe Leserinnen und ...</b></p>
    <p>beim <i>Google</i>-Vergleich gibt ...</p>
    <p>Auf nationaler Ebene ...</p>
    <p>Ausführliche Informationen ...</p>
    <p>Mit freundlichen Grüßen</p>
    <p>Rainer Just<horiz-br/>Dr. Robert ...</p>
    <p>Geschäftsführende Vorstände</p>
  </textual-content>
  <textual-content id="2" style-id="2">
    <h>Aktuelle Entwicklungen ...</h>
    <p>Beim <i>Google</i>-Vergleich ist ...</p>
    <p>Was bedeutet diese überraschende ...</p>
  </textual-content>
</document>
<page>
  <vertical-section id="1">
    <unknown-graphical-content/>
  </vertical-section>
  <vertical-section id="2">
    <horizontal-column id="1">
      <rect-container type="shaded-box">
        <textual-content id="1"/>
      </rect-container>
      <textual-content id="2">
    </horizontal-column>
    <horizontal-column id="2">
      <textual-content id="2">
    </horizontal-column>
  </vertical-section>
</page>
```

Machine learning

Typically used to classify blocks e.g. as individual characters.
Limited ability to recognize larger structures.
“Black box” with confidence measures.



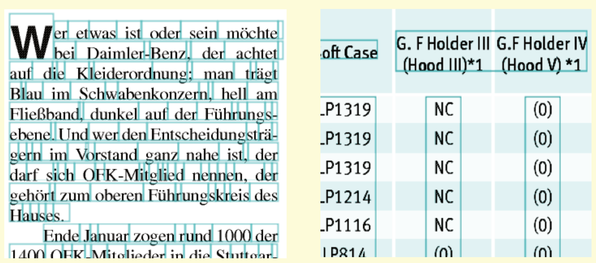
Generative AI

Able to generate plausible answers to textual prompts, such as describing images, extracting articles and converting to HTML.
Very powerful, but may hallucinate.
No confidence measures; result needs to be verified by other means.

Sources: Projection profile image: Wahl et al.: Block Segmentation and Text Extraction in Mixed Text/Image Documents, CGIP 20 (4), pp. 375–390;
Neural network image: Wikimedia Commons

Explainable low-level methods

Simple, rule-based methods may fail to see the “big picture”



Wer etwas ist oder sein möchte bei Daimler-Benz, der achtet auf die Kleiderordnung: man trägt Blau im Schwabenkonzern, hell am Fließband, dunkel auf der Führungsebene. Und wer den Entscheidungsträgern im Vorstand ganz nahe ist, der darf sich OFK-Mitglied nennen, der gehört zum oberen Führungskreis des Hauses.

Ende Januar zogen rund 1000 der 1400 OFK-Mitglieder in die Stuttgar-

oft Case	G. F Holder III (Hood III)*1	G.F Holder IV (Hood V) *1
-P1319	NC	(0)
-P1319	NC	(0)
-P1319	NC	(0)
-P1214	NC	(0)
-P1116	NC	(0)
LP814	(0)	(0)

PDF object-based

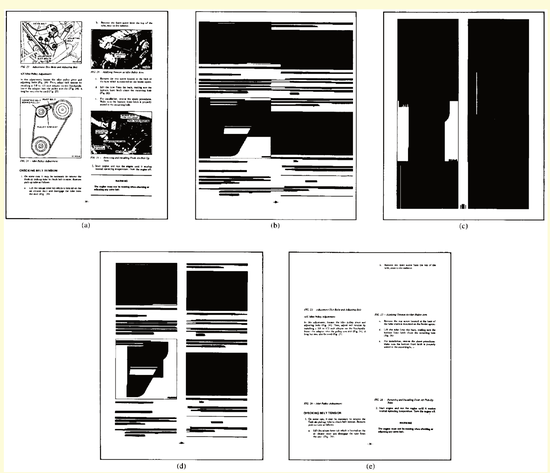
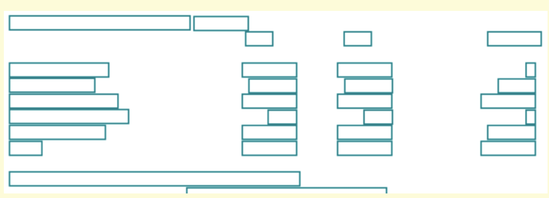


Image-based

No domain knowledge



Explainable AI search

Guided search visits all plausible potential results, which are evaluated according to the model.
High recognition accuracy on inputs that can be correctly represented by the model.