OSDOcr Modules

Gonçalo Afonso

November 7, 2023

1 Introduction

In this document, we will formalize the Old Structured Document OCR (OSDOcr) modules.

2 OCR box Module

2.1 OCR Box

A OCR box represents a container element for a region in a document. Each container may include other containers of lower levels, with the lowest being a word container. Based on a n-ary tree structure.

- **level** : text level of the box. $\{1 : \text{page}, 2 : \text{block}, 3 : \text{paragraph}, 4 : \text{line}, 5 : \text{word}\}$
- **page_num** : only meaningful when multiple pages are processed.
- **block_num** : block identifier in which box is inserted
- **line_num** : line identifier in which box is inserted
- **word_num** : word identifier (applicable if level is word)
- **box** : instance of box class (stores coordinates of bounding box)
- **text** : text recognized inside the box
- **conf** : level of confidence in the text
- **id** : box identifier
- **type** : type of box. ['delimiter', 'image', 'text']
- **children** : children boxes (all of lower level and contained within itself)
- **parent** : parent box (box of higher level that contains it)

2.2 Methods

- **is_empty** : $OCR\_Box \rightarrow Bool$

  Checks if a box container is empty. Every box of level 5 (word) within it has to be empty for a positive result.

- **is_delimiter** : $OCR\_Box \rightarrow Bool$

  Checks if a box group is a delimiter. A delimiter is an empty box container that follows the rule:

$$box.width \geq box.height \times 4 \vee box.height \geq box.width \times 4 \tag{1}$$

  where *box* is the OCR box's Box instance.

- **get_id** : $(OCR\_Box, id : Str, level : Int) \rightarrow OCR\_Box$

  Finds a box container, within higher level box, or itself. The box container is identified by the *id* and the *level*.

- **calculate_mean_height** : $OCR\_Box \rightarrow Float$

  Calculates the mean height of a box group.

- **is_text_size** :
  $(OCR\_Box, text\_size : Float, mean\_height : Float?, range : Float) \rightarrow Float$

  Checks if a box is of a text size. A bpx is of text size if the mean height of the box group is within the range of the text size. Range is by default 0.3.

- **get_delimiters** :
  $(OCR\_Box, search_area : Box, orientation : Str, conf : Int) \rightarrow [OCR\_Box]$

  Gets the delimiter boxes in a box group. The delimiter blocks are the blocks that are delimiters and are inside the search area and respect the given orientation.

3 Engine Module

tesseract_search_img : $img\_path : Str \rightarrow Dict$

Searches text in an image using tesseract. The result is a dictionary with bounding boxes.

tesseract_convert_to_ocrbox : $Dict \rightarrow OCR\_Box$

Turns a dictionary of tesseract results into a OCR box instance.

4 OCR Analysis Module

analyze_text : $OCR\_Box \rightarrow Dict$

Analyzes a box group. The analysis result returns the value of *normal_text_size*, *normal_text_gap*, *number_lines*, *number_columns* and *columns*.

draw_bounding_boxes :
$(OCR\_Box, image\_path : Str, draw\_levels : [Int], id : Bool) \rightarrow img : MatLike$

Draws bounding boxes in an image. The image is loaded from *image_path* and the bounding boxes are drawn in the image according with boxes group given and the levels in *draw_levels*. If *id* is true, the id of each box is also drawn in the image.

estimate_journal_header : $(OCR\_Box, image\_info : Dict) \rightarrow Box$

Estimates the journal header using its box group. The header is estimated by finding the blocks that are delimiters and follow the rule:

$$delimiter['bottom'] \geq image\_info['bottom'] \times 0.5 \wedge delimiter['width'] \geq image\_info['width'] \times 0.3 \quad (2)$$

estimate_journal_columns :
$(OCR\_Box, image\_info : Dict, header : Box?, footer : Box?) \rightarrow [Box]$

Estimates the journal columns using its box group. The columns are estimated by finding the blocks that are vertical delimiters and are within the area between the header and the footer if they exist (ortherwise within the page).

estimate_journal_template : $(OCR\_Box, image\_info : Dict) \rightarrow Dict$

Estimates the journal template using its box group. Returns a dictionary with the header and the columns.

5 OCR Box Fix Module

improve_bounds : $OCR\_Box \rightarrow OCR\_Box$

Improves the bounds of a box group. Not yet finished.

block_box_fix : $OCR\_Box \rightarrow OCR\_Box$

Fixes the blocks boxes in box group. Eliminates empty, non delimiter boxes and eliminates intersections.

join_aligned_delimiters : $(delimiters : [OCR\_Box], orientation : Str) \rightarrow [OCR\_Box]$

Joins aligned delimiters. The delimiters are aligned if they have the same horizontal or vertical value within a range (*is_aligned* for further reading), depending on the orientation.

6 Information Extraction Module

simple_article_extraction_page : $OCR\_Box \rightarrow [Article]$

Extracts articles from a page. Not yet finished.

7 Output Converter Module

boxes_to_text : $OCR\_Box \rightarrow Str$

Converts a box group into a string. The string is the concatenation of the text of each box in the group.