A comprehensive survey of mostly textual document segmentation algorithms since 2008 (2017)

A Resume

November 27, 2018

Abstract

Highlighting the variety of the approaches that have been proposed for document image segmentation

1 Introduction

Document segmentation: dividing the document image into meaningful parts; glyphs, words, text lines, paragraphs, regions. To be used for further content extraction such as text recognition, to determine the reading order, or to classify the document.
More formally for the value of each input element $I(x)$ a segmentation algorithm associates a region number or a label $J(x)$ where x is the element index. x can be the page number, the node index, or even the pixel coordinates in an image $I(x) \xrightarrow{Segmentation} J(x)$

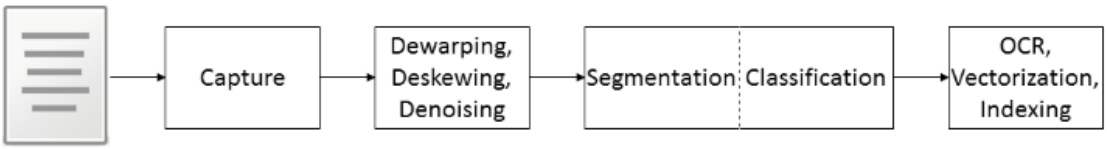


Figure 1: A classical document content extraction process One

There are two kinds of identification: the physical layout and the logical layout. The *physical layout* relates to the nature of the content such as text, typewritten text, graphics, diagram, picture, decoration, etc. The *logical layout* relates to the function of the content such as header, footnote, main body, etc.

2 Topology of segmentation algorithms

Document image segmentation algorithms are typ- ically classified into three groups: top-down, bottom-up and hybrid algorithms.

Top-down algorithms start from the whole page and try to partition it. *Bottom up* algorithms start from a small scale and try to agglomerate the smaller elements (pixels, connected components and patches which is a userdefined scale) into the scale of the whole document.

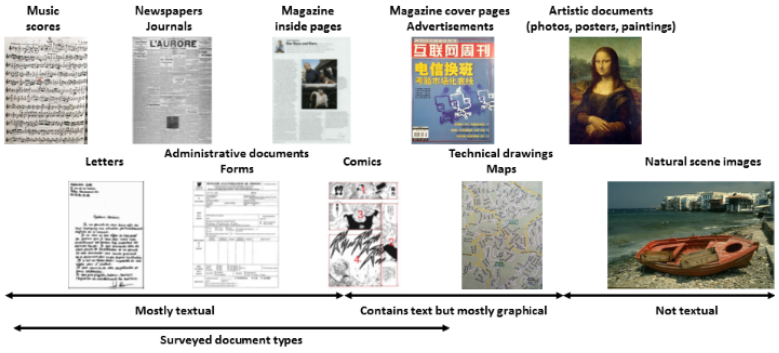


Figure 2: Document typology

Classifying the segmentation algorithms into three groups based on ability to segment any layout:
- **Groupe 1:** the limitation coming from the way the algorithm works (e.g., X-Y cuts for Manhathan layout).
- **Groupe 2:** from the parameters given to the algorithm (e.g., Voronoi requiring different parameters, font size, cc(connected components) size ...).
- **Groupe 3:** Attemps to overcome these limitations (e.g., neural networks).

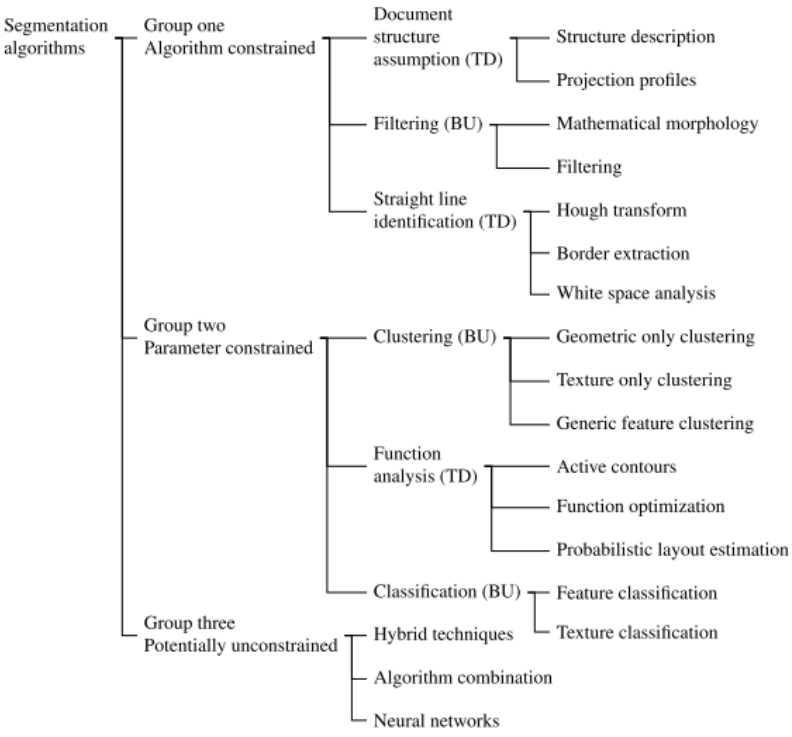


Figure 3: Document typology

3 Goupe one algorithms: Layout constrained by the algorithm

Were first to appear, they aim to segmenting a specific predefined layout such as Manhatann, and they can be used with out any training.

There are three main subcategories in this group:

- The algorithms that make clear assumptions about the document layout,
- The algorithms that use filtering techniques,
- The algorithms that try to identify straight lines.

3.1 Segmentation based on document structure assumption

Segmentation based on document structure assumption They are made for a very specific type of layout hence they are only applicable to documents with a structured layout. This drawback is counterbalanced by their success rate in segmenting this specific layout in comparison with other more flexible techniques and their speed.

Grammar Such as DMOS [1], a layout grammar language, This language can describe any layout and the associated parser recognizes this layout in an image with the association of a label to ech region. It was improved by adding a multiresolution approach with made it flexible enough to sengment handwritten letters.

Side note: DMOS (Description and MOdification of Segmentation) The DMOS method is made of three blocks:
- *Visual clues* are used as input; it can take as input for document analysis and recognition many kinds of visual clues. Basically, it uses primitives extracted at various levels of resolution of the image: connected components, line segments, text lines or any output of more elaborate structure recognition systems (OCR). To combine these inputs, DMOS use perceptive layers, each one is a structure that contains primitives, each layer gives a point of view of the document.
- *Knowledge description* The knowledge description if realized using a specific language called EPF (Enhanced Position Formalism). This rule based language enables to express knowledge on the physical content, but also the logical organization of the documents. The role of the programmer is to describe the best strategy of recognition, that efficiently combines the visual clues coming from the different points of view of the document.
- *Compilation step* Once the EPF description of the kind of document is available, the associated recognition system is automatically produced by a compilation step.

Projection profiles There They make a paving of the document with rectangles. Then they compute the projection profile of each rectangle along several directions. The direction with the highest maximum of Wigner-Ville distribution is that of the text. Then, they use heuristics combined with local projection profiles to detect regions with non homogenous text orientation and text lines.

3.2 Segmentation based on filtering algorithms

Mathematical morphology Such as the usage of a combinations of erosion and dilation to efficiently identify successively the pictures, the graphics and the text. While being a basic type of processing it proves very efficient for the task of document retrieval.
Another apporach is replacing the logical AND of RLSA by an OR, making the algorithm more computationally efficient. Also the Bloombergs segmentation algorithm uses a hit-miss morphological transform to merge broken horizental and vertical lines and a second step to fill holes.

Side note: Hit-and-Miss Transform
The hit-and-miss transform is a general binary morphological operation that can be used to look for particular patterns of foreground and background pixels in an image.
The structuring element used in the hit-and-miss is a slight extension to the type that has been introduced for erosion and dilation, in that it can contain both foreground and background pixels, rather than just foreground pixels, i.e. both ones and zeros. Note that the simpler type of structuring element used with erosion and dilation is often depicted containing both ones and zeros as well, but in that case the zeros really stand for 'don't care's'.
Example : Four structuring elements used for corner finding in binary images using the hit-and-miss transform. Note that they are really all the same element, but rotated by different amounts.:

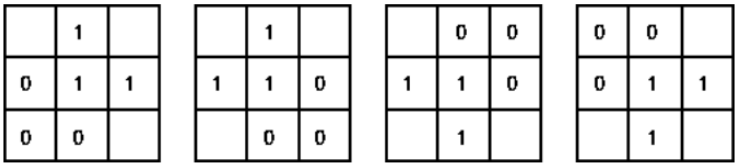


Figure 4: Document typology

After obtaining the locations of corners in each orientation, We can then simply OR all these images together to get the final result showing the locations of all right angle convex corners in any orientation. Figure 3 shows the effect of this corner detection on a simple binary image.

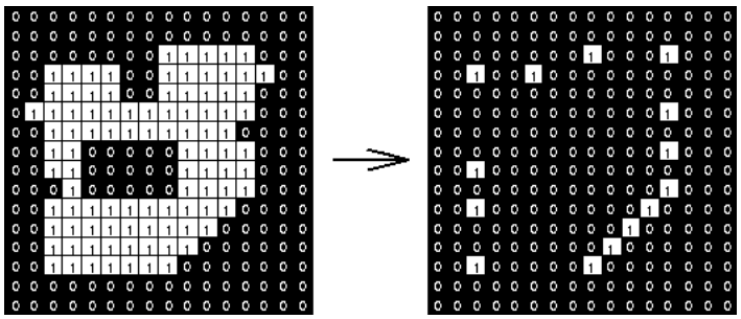


Figure 5: Document typology

3.3 Filtering

The method is based on steerable filters (filters that can be rotated) to detect text lines along five orientations. A heuristic post processing is used to solve the issue of connected components spanning several lines.

3.4 Segmentation based on straight line identification algorithms Three

Based on straight line identification, such as slplit connected components horizontally into blocks based on the average character height. Once this partitioning is done, they apply a Hough transform on the centers of gravity of each block to detect text lines. Another attempts to reconstruct the border of the frames in comic books in order to segment them. They separate the background then they use the Douglas-Peucker algorithm to fit quadrangles onto the candidate frames. Followed by a classification of the frame complexity.

Side note: Douglas-Peucker algorithm The input of the algorithm is a curve represented by an ordered set of points $(P_1, ..., P_n)$ and the threshold $\epsilon > 0$. The output is the curve represented by the subset of the input set of points.
On the first step of the algorithm we search for the farthest point (P_z) from the line segment between the start and the end points $(P_1 and P_n)$. If that point is closer than the threshold (ϵ) all the points between P1 and Pn are discarded. Otherwise the P_z is included in the resulting set. Then we repeat the same step recursively with the right and the left parts of the curve (from P_1 to P_z and from P_z to P_n). Then we merge the results of processing the left and the right parts. Algorithm repeats until all the points are handled.

4 Goupe two algorithms: Layout constrained by the parameters

They try to adapt to local variations in the document in order to be able to segment a broader range of layouts with the same algorithm, by using flexible algorithms with a higher number of parameters which are difficult to train. Thus being limited with the assigned parameters.

There are four main subcategories in this group:

- clustering algorithms based on geometric or texture or a more general set of features,
- The algorithms based on function analysis relying on function optimization
- The classification algorithms trained to recognize the different types of elements based on a given set of features

4.1 Segmentation based on clustering

This is the most popular type of algorithms.

Geometric only clustering *Black and white algorithms.* Liu et al. [2] use a Gaussian Mixture model to classify 310 connected component triplets as text or non text. They use three geometric features (distance, area, density) and thus have trivariate Gaussian distributions. The first order neighborhood of a connected component is computed with the Delaunay triangulation and they use the second order neighbors to obtain all the possible triplets. They also use a specific training called MMS which maximizes the class separability. Although the 315 algorithm is not made for color images, it is tested on binarized color advertisements and magazine cover pages (see github page for Delaunay triangulation).
Another approach is an improved Voronoi algorithm which adapts the Voronoi algorithm to the local spatial context, an other version with fuzzy edges called CVS was introduced.
Other algorithms were introduced: metric learning based on geometric features, geometric clustering to segment horizental and vertical text, using parallel line regression.
Other algorithms. such as using nearly uniform colors to segment the image componenets with a pixel clustering based on neighborhood RGB color distance.

Texture only clustering They highlight the importance of a multi-resolution approach to reduce the noise in pixel clustering techniques. Working at pixel level allows the clustering of many different types of objects such as drop caps, a specific kind of graphic, text, text fonts, etc.

Generic feature clustering Six Such as cutting the document into blocks which are then multi-thresholded to create several layers. The connected components of each layer are identified and grouped across blocks based on a predefined set of features.

4.2 Segmentation based on function analysis

They have the advantage that, based on the flexibility of the functions, one can select how much they will follow the contours of the elements to segment. This can be helpful if we want to have a rough outline of the document regions or if we want to segment precise elements such as warped text lines.

Active contours It works by adding coupled snakelets (a kind of non closed active contour) on the top and bottom of a connected component and by deforming them based on the vertical component of the gradient vector flow. The snakelets are then extended laterally in order to include neighboring connected components.

Function optimization They usually define a cost or energy function which needs to be minimized.

Probabilistic layout estimation Such as performing an estimation of the number of text lines with a blur filter and then use a variational Bayes approach to segment the image rescaled at 75 dpi. This improves slightly the state of the art on a large but not very challenging data set.

4.3 Segmentation based on classification This

This is the second most popular type of algorithms with 30 algorithms. A noticeable difference in the scientific work when compared with the clustering is the fact that classification algorithms all require training.

Texture classification Such as using three Dynamic Multi-Layer Perceptrons (DMLP) at three resolutions to segment historical documents. Each DMLP uses the label output of the DMLP at a lower resolution plus texture features at its resolution. Each level processes only part of the labels produced by the lower level in order to refine these specific labels.

Feature classification *Black and white algorithms.* Such as automatically selecting the appropriate features based on the desired typewritten text classification accuracy. It selects the first 100 fea- ture vectors of a Principal Component Analysis (PCA) of the character images. Then any new text is projected into this new space.

Gray-level algorithms. After extracting candidate word blobs with projection profiles, they introduce Gradient Shape Features (GSF) to refine the segmentation and classify the text as handwritten or typewritten with a support vector machine (SVM).

Color algorithms. Such as using an SVM to classify Gabor and edge features followed by a CRF to include the local spatial context. The the CRF improves the performance by 2 points. Or using other methods such as MLP and GMM, Wei et al.[5] compare the performance of SVM, MLP, GMM classifiers ans find that SVM ans MLP outperform GMM bu cannot conclude which one is best.

5 Goupe three algorithms: Layout potentially unconstrained

These algorithms try to overcome the shortcoming of the others by hybridizing them, combining them or with advanced neural networks.

There are three main subcategories in this group:

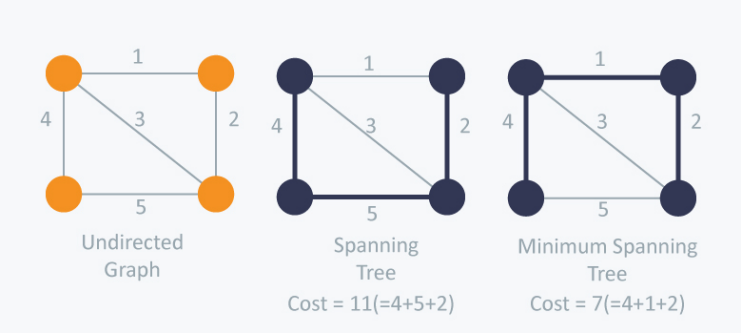
- The hybrid algorithms combining several other algorithms in symbiosis,
- The combination algorithm combines the results of several algorithms to effectively improve them.
- The neural network algorithms make use of artificial intelligence to automatically learn significant features and perform the required task.

5.1 Segmentation based on hybrid techniques

One of the methods is based extracts text lines from complexe documents with multi-oriented text in several languages, using MSER to extract candidate characters which are then classified as text of non text with a fast Adaboost classifier, the low confidenc text is further evaluated with a CNN, the next step us a coarse line extraction with geometrical grouping based on a linearity constraint, this graph is refined by a minimum spanning tree. The last step is an energy minimization refinement of the lines.

Side note: Maximally Stable Extremal Regions MSER is a method for blob detection in images. The MSER algorithm extracts from an image a number of covariant regions, called MSERs: an MSER is a stable connected component of some gray-level sets of the image.
MSER regions are connected areas characterized by almost uniform intensity, surrounded by contrasting background. They are constructed through a process of trying multiple thresholds and the selected regions are those that maintain unchanged shapes over a large set of thresholds.

Side note: Spanning Tree Given an undirected and connected graph $G = (V, E)$, a spanning tree of the graph G is a tree that spans G (that is, it includes every vertex of G) and is a subgraph of G (every edge in G the tree belongs to).
Minimum Spanning Tree?
The cost of the spanning tree is the sum of the weights of all the edges in the tree. There can be many spanning trees. Minimum spanning tree is the spanning tree where the cost is minimum among all the spanning trees. There also can be many minimum spanning trees.
Minimum spanning tree has direct application in the design of networks. It is used in algorithms approximating the travelling salesman problem, multi-terminal minimum cut problem and minimum-cost weighted perfect matching. Other practical applications are:
- Cluster Analysis,
- Handwriting recognition,
- Image segmentation.



5.2 Segmentation based on algorithm combination

A procedure to significantly improve the performance of individual segmentation algorithms by combining their results. The procedure is based on the overlap of the regions produced by the algorithms. They are considered as good above 90% overlap. They use the regions that have more than 70% to compute the values of a task based set of features. All regions below 90% undergo a splitting based on their intersection followed by a merging starting from the regions with the highest overlap. This a procedure to significantly improve the performance of individual segmentation algorithms by combining their results. The procedure is based on the overlap of the regions produced by the algorithms. They are considered as good above 90% overlap. They use the regions that have more than 70% to compute the values of a task based set of features. All regions below 90% undergo a splitting based on their 545 intersection followed by a merging starting from the regions with the highest overlap. This

5.3 Segmentation based on algorithm combination

Methods such as bidirectional long-short term memory neural network (BLSTM) to detect text lines and paragraphs, They tackle the issue of modeling gaps and interlines and conclude that each should have its own class. They also show that using specialized neural networks trained on a specific set is better than using a single system trained on a more varied data set.

6 Benchmarks

- MADCAT program by DARPA : LINK
- RIMES database by A2iA : LINK
- MAURDOR by DGA : LINK
- ICDAR competitions : LINK

7 Datasets

IAM Historical Document Database LINK & MAURDOR

8 Metrics

Most evaluations are based on the same principles of counting: false alarms (adding a region), misses (removing a region), merges (two or more regions in one), splits (one region in two or more) and matches (properly segmented region). Algorithms that tend to merge (respectively split) regions are said to under-segment (respectively over-segment) the documents.

9 Main conferences / journals

- ICDAR : International Conference on Document Analysis and Recognition,
- DAS : IAPR International Workshop on Document Analysis Systems,
- PR : Pattern recognition,
- IJDAR : International Journal on Document Analysis and Recognition,
- DocEng : ACM Symposium on Document Engineering (DocEng),
- DRR : DocPattern Recognition Letters (PRL)ument Recognition and Retrieval,
- ICPR : International Conference on Pattern Recognition.

References

[1] Bertrand Couasnon, *DMOS, a generic document recognition method: application to table structure analysis in a general and in a specific way*

[2] L. Liu, H. Fu, Y. Jia, *Gaussian mixture modeling and learning of neighboring characters for multilingual text extraction in images*

[3] Mudit Agrawal; David Doermann, *Voronoi++: A Dynamic Page Segmentation Approach Based on Voronoi and Docstrum Features*

[4] Mudit Agrawal; David Doermann, *Context-aware and content-based dynamic Voronoi page segmentation*

[5] Wei, M. Baechler, F. Slimane, R. Ingold, *Evaluation of SVM, MLP and GMM classifiers for layout analysis of historical documents*