A Hierarchical Clustering Method Aimed at Document Layout Understanding and Analysis

Costin-Anton Boiangiu, Dan-Cristian Cananau, Bogdan Raducanu and Ion Bucur

Abstract—This paper presents a new approach towards creating a type of hierarchy for document image page using the information given by the Delaunay triangulation. The steps of the algorithm are presented under the form of a cluster tree containing the information of the page in structures such as collections of pixels and using the distance between them as a binding measurement. The final result provides the page segmentation into clusters containing pictures, titles and paragraphs.

Keywords—cluster tree, contour detection, Delaunay triangulation, page hierarchy, pixel entities.

I. INTRODUCTION

The development in the area of scanning and printing devices has known a great expansion in the last years. And because of this reason there have been further increases in the expectations of the document content recognition and conversion. The purpose is the expansion of the electronic interpretation of the document by understanding the logical structure (chapter delimitation and titles, sections, headings, paragraphs, authors and affiliation, annotation, footnotes, references, commentaries, related pictures and schemas, page number) [13]-[15].

The goal of this paper is to present a solution towards determining this layout and to create a form of hierarchy for the document using this layout and the first step is to find the basic entities in a document and with them to create such a structure. These basic entities are represented by the separators, which can be roughly classified based on their shape or geometrical characteristics into:

- Line separators;
- Line-based separators;
- White space separators;
- Arbitrary-form separators;

The common knowledge on separators presents them as image segments that have certain geometrical characteristics, like, for example, in a horizontal line the width is much greater than the height. Most algorithms use only this information towards detecting such entities and more evolved approaches respect the angle orientation of the separators for broken line detection. Such approaches are shape dependent and take into consideration just line separators. Better ones use the concept of distance and provide a mathematical solution for the detection like in the examples found in [11], [12], [23].

For the white-space detection, most algorithms are somehow similar to the ones used for lines because the detection is based on the fact that the number of white pixels found on a direction is greater than the number of the pixels found on a direction orthogonal to the initial one. Even though this approach has the same disadvantages as the one used for lines because of the size and orientation dependency, it proves to have a greater degree of certainty. However none of this type of approaches is satisfactory and a geometrical independent method is required for correct detection of separators (for further line detection algorithms refer to [2]).

In this paper a reliable approach will be presented, approach based on creating a hierarchical clustering structure [3].

What differentiates this method from others presented in similar papers is its type. It uses a “top-down” one instead of a “bottom-up”, which means that it does not have the purpose of grouping different objects into collections, but instead it breaks the collections into objects. The Delaunay triangulation ([8]-[10], [24]) presents the perfect mathematical tool towards obtaining neighborhood relations and further using them to simulate the characteristic of the human eye of “connecting” similar elements.

The final structure will be presented as a cluster tree. This will combine the results obtained from the triangulation a specific cluster tree construction algorithm. By using such a structure entities will be gathered into single components based on the distances computed by the triangulation. The tree will use the Euclidian distance as its measurement and will introduce a new definition, the “hierarchy distance”, in order to facilitate the merging operations done on the entities. All of these aspects shall be presented in the following pages.

II. PROBLEM SOLUTION

There are several steps that have to be followed in the correct order to obtain the final tree hierarchy. The first steps have been presented in a previous article ([4]) and are presented succinctly because they are mandatory for the correct completion of the final step.
A. Preprocessing

The initial step is a preprocessing one, because the input has to be prepared for the requirements of the algorithm [19]. One of the most important aspects of this approach is that it uses a black and white document and in order to achieve this goal for every document a simple black and white conversion has to be made regardless of the initial color pattern. There are several algorithms that serve this purpose we have selected the most suitable one for our kind of input documents [1],[21].

B. Contour generation

Next, the input selection has to be done and this implies generating the image segments (further referred as entities of the image). A collection of connected black pixels represents an entity, which is easily determined with the help of a simple algorithm that starts from a black pixel and passes through all neighboring black pixels until there are only white neighbors [18].

By repeating this algorithm for all non-visited black pixels the entities are obtained in the end. There are several shapes that can bind a collection of black pixels. The actual bounding shape is in fact a polygon which contains and approximates the entity or collections of entities and in Fig. 2 we present the most common one: the rectangle.

For the presented approach the bounding rectangle is not used, but instead a contour of the current entity is taken into consideration. Because each entity can be seen as a collection of horizontal segments, the contour is generated from the extremities of each such segment of the entity, with the mention that all the extremity points of the segments which cannot be seen directly from an external point of view are not taken into consideration.

A simple example of this algorithm is presented below in Fig. 3. Another type of contour generation algorithm is presented in [5], [17].

C. Delaunay triangulation

After the actual contour selection the next step is the use of the constrained Delaunay triangulation algorithm. In this way all the entities will be connected to each other. However, this is more then we need and so a processing of the obtained Delaunay triangles has to be done. All the triangles that connect more or less than two entities are eliminated and the final result reveals only entities connected in groups of two.

This fact allows the creation of two types of points, which are named as a convention in this paper: current and destination points. The names come from their characteristic of belonging or not to an entity.

By using the Delaunay triangulation each entity has several triangles starting from it and going towards another entity. The points of the triangles which are on the current entity are the current points and the ones that belong to the triangles, but are situated on another entity different from the current one are called destination points.

D. Proximity generation

The proximity is an “entity to entity” relation. The proximities are generated by iterating toward the triangles contained in the constrained Delaunay triangulation and
filtering triangles that join two different (inter-triangles) entities. Triangles that are generated inside one entity (intra-triangles) or between three distinct entities are discarded from processing.

The proximity structure holds vital statistics regarding the entity-to-entity relation like: the pair of entities, the minimum square distance inside Delaunay inter-triangles, the number of connections points in both entities, the area of connection and other measures that may be relevant depending on the processing type.

E. Separators

There are several classifications of separators based on their geometrical form or characteristics as stated in the introduction, but all of them have one important thing in common which puts them into the spotlight. By using the already presented Delaunay triangulation and detecting the current and destination points a statistics can be made based on their ratio.

The result reveals a very important characteristic of separators: they have far more current points than destination points because they extend to several entities in size independent of the orientation or angle. In this way the separators are detected and a line can be drawn between them and regular characters, like letters or punctuation signs.

The next step is to use this information and introduce it in a hierarchy of the page. By doing so, we will get the text areas which will be bounded by page edges or separators inside the hierarchy tree.

III. CLUSTER TREE

Our method creates a hierarchical model of the input entities by building a special type of multi-way tree called a cluster tree. The entities will become leafs in such a tree and the internal nodes of the tree represent clusters of entities. The diameter of a cluster is the maximum distance between any two entities belonging to that cluster or between any adjacent entities that may form a chain to “connect” any entity pair (Fig. 4). The purpose of this tree is to group the entities into clusters with diameters in increasing order of magnitude. Thus, the root of the tree corresponds to a cluster with the largest possible diameter (if this cluster would represent the entire page, then its children would represent top level elements like paragraphs or images).

This hierarchical model is used in collaboration with the separator information obtained at the previous step to build the layout of the page [6]-[9].

There are two courses of action that can be considered when discussing the design of the hierarchy tree. The first one is to use as input the extreme points and the Delaunay triangulation. The extreme points are the points on the contour of the entities.

The tree construction algorithm starts by computing for each pair of entities the minimum length Delaunay triangle edge that connects them. The algorithm constructs the tree in a bottom-up fashion. It starts with a random entity and builds a cluster around it. It will first find the closest entity to this initial entity and add it to the cluster. Next, it finds the closest entity to either of the two and if the distance to this entity is of the same order of magnitude as the distance between the first two, the third entity will also be added to the cluster. Similarly, the algorithm will continue to add entities until the closest entity is of a bigger order of magnitude and thus, cannot be part of this first cluster. The rest of the clusters are constructed in the same way, with the exception that the algorithm now may also add the closest cluster and not just the closest entity.

When the algorithm ends, it produces the desired tree model which accurately describes the hierarchy of the page.

The second approach to constructing the hierarchy is to use points from the bounding shapes of the entities and the distance between the bounding shapes as a metric.

A good choice for the bounding shape is the convex hull. The idea in this case is to compute the convex hull of each entity based on its contour points and based on this result to compute the minimum distance between bounding shapes and use this, as before, as the minimum distance between the two entities. From now on the algorithm is exactly like the one presented above. It begins by constructing a cluster from an empty set by growing it with the closest entity, in the sense of the minimum distance between the bounding shapes. The algorithm continues to build the other clusters in the same manner, just that now, a cluster can also contain other clusters, if the distance between them is the same order of magnitude. The algorithm finishes and produces a hierarchical model of the page, where the root of the tree represents the entire page; its children represent high level layout elements like paragraphs, images, tables, titles, headings, etc. Their children represent smaller elements like text lines, graphical lines, and so on. The leaves represent the smallest elements, which commonly are characters.

![Fig. 4 – simple cluster tree: internal nodes are labeled with the cluster diameters.](image-url)
ennoch meinte Rezoldt: „Es liegt hinter uns.“ CDU-Politiker Hinze sagte: „Die Beziehungen“

derungen auf dem Ar-

en laut Rezoldt „nur ge,

wenn die Koalition ihre

Das bedeutet mehr Wett-

Staat, vereinfachte Pla-

nungsverfahren, groß-

und verbesserte Risikoo-

g. „Eine wirtschaftspoliti-

cwärte – so wie von der

chaft keine Arbeitsplat-

le: die „Fahrte ins rot-gri-

räume einen „Kataskro-

e Menschen in Deutsch-

en den Jahreswirtschafts-

men des Stammes und

keit“. Die möglichen

Turbulenzen in Südöst-

chung der Konjunktur in

e Zinsanpassungen im

mit der Europäischen

und die schwache Bin-

zonen von der Bundesre-

gegessiert, anstatt wirt-

August vorzugehen“, kri-

gierte Deutsche Fraktions-

fuchs und der Wirt-

nicht Schwachheit.


diker benannten die

der Regierung als ge-

scheint. Sie habe zu immer neuen Re-

korden bei der Arbeitslosigkeit, bei der

staatlichen Verwaltung, bei der Steuer-

Abgabenbelastung sowie bei Preisen ge-

führt. „Da nützt auch die Wahlkampfhilfe

Spitzensprecher aus den Wirt-

schaftsverbänden nichts“, so die SPD-

geordneten.

Auch die Deutsche Angestellten-

werkshaft (DAf) und der Deutsche Ge-

werkshaftsbund (DGB) bewiesen die

volkasungsierend. Die Arbeitso-

igkeit werde von der Regierung

onsen, unverbindliche Parteisprechen, dem

s auf der Spitze der Regierung.

BWV-Hauptgeschäftsführer Dieter Här-

wurde von der FDP, die Gewerkschaften

politischen Lager offen zu unterstützen.

Die Wirtschaftsverbände seien

schließlich kein „Räumeralarmvertrag“.

Berichte Seite 4 und Wirtschaftsteil

Öffentlicher Dienst

Schlichter zeigen sich

„gemäßigt optimistisch“

igt BREMEN. 11. März. Die beiden

vorsitzenden der Schlichtungskommission

für die schwierige Tarifverhandlungen im Öf-

fentlichen Dienst sind nach eigenen Aus-

sagen „gemäßigt optimistisch“, da sie ein

Kompromisspaket schnüren könnten. Die

Tarifparteien seien einig geworden, sag-

ten Bremsens Staatsratsherr Hans Koch-

schneck (SPD) und der frühere Regie-

rungsbevollmächtigte. Der „gemeinsame

Tarifvertrag“ der freien Bundesländer hat

aus dem Tage der Tarifgeschäftssprü-

che in Bremen. Die Verhandlungen sollen

am Montag fortgesetzt werden.

Die „allgemeine verfolgte Linie“ laut

Angaben der Schlichter darauf hin,

is und Gehälter für die 3,2

Millionen Angestellte und Arbeiter von

Bund, Ländern und Gemeinden nur ge-

ring zu erhöhen, um Arbeitsplätze zu si-

chern. Zudem sollen die Obi-Stiften im West-Niveau angeschlagen werden. Ko-

schneck und Wagner setzen sich auch da-

für ein, steigende Kosten der Alters-Zu-

satzversorgung abzuwenden. Diese wird

bisher nur von den Arbeitgebern finan-

ziert. Nach der Ansicht sollten auch die

Beschäftigten Beiträge zahlen.

Die Arbeitgeber haben ein Gegen-

Lohn- und Gehaltserhöhung angeboten,

aß jedoch festhielt, dass die Gewerkschaften

ahmen. Deren Forderungen würden 4,5

Prozent Mehrausgaben bedeuten.
Fig. 5 and Fig. 6 represent two versions of the same image before and after the Delaunay triangulation. The first image contains the unaltered document page, while the second one is the result of applied Delaunay after taking into account the bounding shape.

As it can be seen, entities are connected one to another by a thick collection of edges, which are all drawn with the same color in order to emphasize the unity of these lines. Each color contains only the connections between the points on the bounding shape of only two entities and so, it provides a good visual measurement of the distances.

Form the entire collection of connections only the smallest distance is selected and taken as input into the algorithm.

A. *Hierarchy Model – Cluster tree*

As described, a cluster tree can be constructed to model the page hierarchy. In this tree all the leaves represent the input data, the entities. The leaves are grouped into clusters; each cluster is represented by a tree node which is labeled with the diameter of the cluster.

The idea of the cluster tree is that any two elements inside a cluster have a distance no more than the cluster diameter. This also means that each sub-tree of the structure is a cluster in which all of the nodes are closer to each other than to any other node outside that cluster. Fig. 7 shows an example of a cluster tree structure.

As it can be seen in the example, the diameters of the clusters increase if traversing the tree from bottom to top. Each node is included in exactly one cluster and it has no children. The labels of the non-leaf nodes seen in the example represent the maximum distances inside each cluster, or the cluster diameter. For example the cluster “ab” composed of the entities “a” and “b” has the maximum distance 20, which means that no entities inside this cluster are more than 20 units apart. Also, as stated in the definition of the cluster tree and as an implicit effect of the construction algorithm, there is no node lying less than 21 units from either “a” or “b”.

In this example, the order of magnitude is thought to be different if the diameters are not equal. Practical implementations use a threshold to establish if two entities can be part of the same cluster.

In the context of layout analysis, when talking about the distance between two entities we shall actually be referring to the diameter of the cluster that those entities are part of. This will be referred to as the *hierarchy distance*, and it is opposed to the *Euclidian distance*. The Euclidian distance is used to build the cluster tree, while the hierarchy distance is used as a layout space measure.

The Euclidian distance is a well-known term which defines the minimal path between two points, the length of the segment that connects them. The hierarchy distance however has a different meaning. In the following example the distance between the points “A” and “D” is 90.

![Fig. 8: the meaning of hierarchy distance](image)

The hierarchy distance between “A” and “D” is 45 and can be obtained from the cluster tree. Because “B” and “C” form a cluster and then join with “A” into another cluster before joining with “D” all three points have the same hierarchy distance to “D”, which is the Euclidian distance between “C” and “D”. And so, this new measurement unit provides a good mean of evaluating cluster closure.

Now that the terms used in this paper have been explained there are several steps that have to be followed in order to obtain the desired clusters. The first step was to create the cluster tree. Next, the information contained in the clusters is used to join different entities or groups of entities depending on hierarchy distances. In the following group of images the creation of the clusters can be observed and the final result of splitting the document in zones with similar characteristics is obtained.
In the end all the zones have been detected properly. The iteration process of joining clusters can continue and the whole page will be seen as a standalone cluster, but this would be too much. The purpose is to create zones of similar information in the page and after a number of iterations that algorithm must stop.

The charts provide an overview of the distance values for the current tested image by plotting the histograms of such values.

The result obtained for the given picture allows the detection of titles, paragraphs and even articles. However without a mechanism for result measurement there is no knowing when to stop the iteration process.

As it can be seen this is the first iteration of the process where only the closest entities were connected into clusters. To have a better view on the clustering, only a small part of the initial picture has been taken for the first set of result images and the clusters have been bounded with rectangles. For a picture of such sizes this has almost no effect and does not provide any aid in splitting the document in zones. The next picture however is taken after only a few more iterations. It can be observed that groups of entities have been connected by the algorithm and some sort of cluster hierarchy has been created.

By continuing the process of joining the clusters the presented paragraph of the initial image has been finally detected as an independent zone.

For this purpose several concepts will be introduced. First of all by joining the entities there is one measure that always changes making each iteration different from all others. By using this knowledge a mechanism for assessing the results and finding the steps in which a relevant change has been made can be developed.
Therefore the most important thing that changes with each iteration and provides relevant information on the clusters is the rectangular area of the clusters.

This can be divided into three different types: total rectangular area (the sum of all the rectangles that bind the clusters; non-overlapping rectangular area (the sum of all the rectangles that result from the intersection of all the rectangles that bind the clusters), non-overlapping rectangular area (the sum of all the rectangles that bind the clusters from which the overlapping area is subtracted). The above charts present the measurement stated above at each iteration.

From these results we can determine the inflexion points, the points in the graph where the function changes its slope...
These points are represented with a white line in the graph. In this case the function is the type of area used for the chart.

By evaluating these results it can be stated that at each inflexion point there has been an important change in the graph. For example, when the next value of the total area is higher than the current one this means that the clusters have been joined together into a bigger one.

Fig. 15: the final result which finds all the important zones of the page inside an individual cluster
However, when the next value is lower than the current one some clusters that were inside a bigger one have been connected to that one and so the area has decreased. By monitoring the changes in slope sign from increase to decrease and decrease to increase it can be observed that the most important changes happen only at those times. And so a decision to stop at a given iteration has to be made taking into accounts only these points. In order to obtain the best cluster hierarchy one of the last such points has to be considered as the stopping point of the algorithm.

In the Fig. 19, the total rectangular area has the value 20, the overlapping rectangular areas has the value 4 and the non-overlapping area has the value 11 because in the given example we assumed that every rectangle has an area equal to 4 units (2 by 2).

**IV. Conclusion**

The approach presented above reveals a good tool for page layout analyze by allowing the selection of different groups of entities. This is done by cutting the tree at different levels and so obtaining the corresponding groups. Such a method allows the correct detection of paragraphs, headlines and other types of layout elements with a simple and easy to implement algorithm that can also have various applications outside the document content conversion area.

By using various mathematical solutions and algorithms together with common knowledge content analyze the correctness of such an approach can be easily proved and verified.

The layout analysis method presented in this paper is a natural development of a hierarchical clustering process. Imagine that you look at a document and, progressively, you slowly move away while continue to look at the document.
What will happen: the image will become kind of “blurry”, you will miss some details of the image, you will not be able to read words from normal text paragraphs but you will still be able to see where the paragraphs, headlines, tables and images are placed and how the document is structured.

Moving farther away of the document will enable you to see less of the document detail but more of the document layout upper-structure. Is something that may be somehow simulated by applying a pyramidal resampling of the image until the image implodes itself in only one dot. This intuitive process is able to see where the paragraphs, headlines, tables and images are placed and how the document is structured.

Furthermore, the human eye is more sensitive to rectangular-like structures and, as a result, the rectangular reconstruction of clusters inside the document is favored through the usage of some cluster-area functions that needs local maximization:
- sum of the rectangular areas of the elements;
- sum of the rectangular areas of the elements, excluding rectangular overlaps;
- sum of the rectangular overlaps of the elements.

Having multiple clustering-measures will enable the clustering algorithm to choose the best approach for the current document layout.

ACKNOWLEDGMENT

We want to thank our colleagues from the University “Politehnica” of Bucharest, to the staff of “Computer Science and Engineering” department, for their ideas, research support and useful advices.

REFERENCES