

A Hierarchical Cluster Tree Approach Leveraging Delaunay Triangulation

Cristian AVATAVULUI ¹

Costin-Anton BOIANGIU ²,

¹ University Politehnica of Bucharest,
060042 Bucharest, Romania, ORCID ID:
0009-0003-4760-4055,
cristian.avatavului@gmail.com

² ORCID ID: 0000-0002-2987-4022,
costin.boiangiu@cs.pub.ro

Abstract: *This research introduces a robust and reliable technique for structuring document image pages hierarchically, harnessing the power of Delaunay triangulation. Central to our approach is the formation of a cluster tree, which encapsulates the page's content through strategically exploiting layout elements arrangements and their relative distances. By applying our technique, we proficiently categorize the page into distinct clusters encompassing images, titles, and paragraphs. The consequent hierarchical framework, founded on the cluster tree, establishes a durable and trustworthy blueprint of the document layout, thereby accelerating document comprehension and examination.*

Keywords: *hierarchical clustering, document image layout analysis, Delaunay triangulation, cluster tree formation, layout element segmentation, advanced document image processing.*

How to cite: Avatavului, C., Boiangiu, C. A. (2023). A Hierarchical Cluster Tree Approach Leveraging Delaunay Triangulation. *BRAIN. Broad Research in Artificial Intelligence and Neuroscience*, 14(3), 408-433.
<https://doi.org/10.18662/brain/14.3/482>

1. Introduction

In recent years, advances in scanning and printing technologies have led to increased expectations for document recognition and conversion. The goal is to expand electronic document interpretation by understanding of the logical framework. This involves grasping elements such as chapter divisions, titles, sections, headings, paragraphs, author details, affiliations, annotations, footnotes, references, comments, associated visuals, diagrams, and page numbering (Nagy et al., 2004; Baird, 2002).

This paper seeks to present a method for determining the document layout and creating a hierarchical structure based on this layout. Identifying the fundamental connected components within a document and utilizing them to build the structure is the initial step. Separators can be broadly classified into different categories, including line separators, separators based on lines, separators based on white space, and separators with arbitrary forms. These distinctions are determined by their shapes or geometric attributes.

Separators are depicted in conventional knowledge as image segments with particular geometric properties. Typically, the breadth of a horizontal line is significantly greater than its height. Most algorithms rely purely on this information to detect these connected components, whereas more sophisticated approaches consider the angular orientation of separators when detecting broken lines. However, these shape-dependent strategies concentrate predominantly on online separators. As demonstrated in, Kise et al. (1998); Wang et al. (1999); and Zhu & Doermann (2007), more sophisticated methods incorporate the notion of distance and offer mathematical approaches for detection.

White-space detection algorithms frequently resemble line detection algorithms, as they rely on the observation that the quantity of white pixels encountered along one axis exceeds the count in the orthogonal axis. Despite having the same limitations as line-based methods due to dimension and orientation dependencies, this method provides a higher level of confidence. However, none of these methods entirely satisfy the requirements, necessitating a geometry-independent approach to accurately detect separators (for additional line detection algorithms, see, Qumsiyeh (1995)).

This paper presents a dependable approach based on the development of a hierarchical clustering structure (Jain & Dubes, 1988). This approach stands apart from the methods introduced in existing literature due to its utilization of a "top-down" strategy instead of a "bottom-up"

approach. This method deconstructs collections into individual objects, as opposed to the preceding one, which concentrates on grouping objects into collections. The Delaunay triangulation (Lee & Schachter, 1980; Lee & Lin 1986; Guibas & Stolfi, 1985) is the optimal mathematical instrument for establishing neighborhood relationships and simulating the human eye's ability to "connect" similar items. The outcome's structure is illustrated as a cluster tree through the combination of the triangulation outcomes with a specific algorithm designed for constructing cluster trees. This structure permits the aggregation of connected components into cohesive components based on triangulation-calculated distances. The tree employs Euclidean distance as a metric and introduces a novel concept called "hierarchy distance" to streamline the merging operations executed on the connected components. The subsequent sections of this paper will elaborate on these points.

2. Previous work

Over an extended period, Document Image Analysis (DIA) has remained a topic of significant scholarly fascination within the domain of computer vision and image processing, owing to its complex challenges and vast applications in digital libraries, character recognition, and automatic data entry (Wenyin & Dori, 1997; Jain & Yu, 1998). Layout analysis, which involves segmenting a document image into its constitutive elements (titles, paragraphs, and images), forms a key aspect of DIA.

Traditional layout analysis techniques have predominantly been rule-based, leveraging heuristic information about the geometry and location of different elements (Baird, 2005). While effective, these methods are often criticized for their inflexibility and sensitivity to noise. More recent works have introduced machine learning-based techniques, which offer more adaptive and robust approaches to layout analysis (Sarkar et al., 2004).

The pioneering introduction of employing clustering methodologies in layout analysis was attributed to Sarkar & Bhowmick (2011). This method involved grouping pixels into clusters to represent individual document elements. Although the methodology demonstrated promising results, it had limitations in adequately representing the hierarchical relationships between different elements.

The mitigation of this limitation was achieved through the introduction of hierarchical clustering methods. Gong & Liu (2016) developed a method that used hierarchical clustering to generate a tree-like structure, known as a cluster tree. This provided a more robust and

representative model of the document layout, effectively encoding the hierarchical relationships between different elements.

The use of geometric algorithms, specifically Voronoi diagrams and Delaunay triangulation, in layout analysis has also been explored. Voronoi diagrams have been employed for text-line detection and separation, exploiting the property that neighboring characters are closer to each other than to characters in other lines (Kise et al., 1998).

In parallel, Coustaty et al. applied Delaunay triangulation in the context of document image layout analysis (Coustaty et al., 2011). The technique has shown effectiveness in distinguishing between different types of elements, primarily due to the geometric properties of Delaunay triangles. However, the incorporation of Delaunay triangulation in generating hierarchical representations of the document layout has not been extensively explored.

Therefore, this paper aims to build upon these studies by proposing a novel methodology for layout analysis that combines hierarchical clustering, Delaunay triangulation, and cluster tree representations to provide a robust and detailed document layout representation.

3. Problem Solution

The paper at hand builds upon our previous work in (Boiangiu et al., 2008), adding some new explanations, better test scenarios, a more adequate dataset, and a new postprocessing phase aimed at further improving the proposed method results.

To attain the final tree hierarchy, a series of sequential steps must be taken. The initial stages, which are necessary for the successful completion of the final stage, have been described in a previous publication (Boiangiu et al. 2008) and will be summarized here.

3.1. Preprocessing

The first step involves preprocessing the input data to fulfill the algorithm's requirements (Otsu, 1979). Important to this strategy is the utilization of black-and-white documents. Therefore, regardless of the initial color scheme, each document is converted to black and white. Multiple algorithms exist for this purpose, and the most suitable one for the input documents has been chosen (Otsu, 1979; Sauvola & Pietikäinen, 2000).

3.2. Contour Generation

Subsequent to input selection, the process involves generating image segments, denoted as "connected components" within the image. A group

of connected black pixels constitutes a connected component, a determination that can be accomplished through a straightforward algorithm. Beginning with a black pixel, the algorithm traverses all black pixels until only white pixels remain as neighbors (Boiangiu et al., 2008).

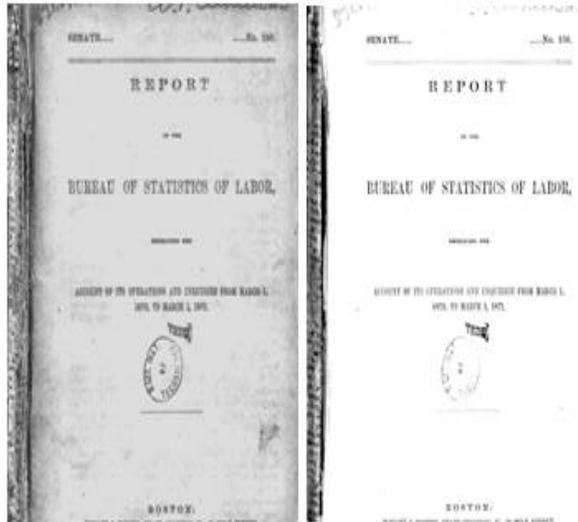


Fig. 1: Conversion of the initial grayscale image into black and white format.
Image sourced from Boiangiu et al. (2008)

The procedure is iteratively applied to all black pixels that have not been visited, resulting in the acquisition of connected components. As shown in Figure 2, a collection of black pixels can be enclosed by a variety of configurations, with the rectangle being the most common. Nevertheless, the presented method does not employ the bounding rectangle. Instead, it evaluates each entity's outline. As each connected component can be interpreted as an assemblage of horizontal segments, the contour is formed by linking the endpoints of these segments. Significantly, endpoint vertices of segments that are not directly visible from an external viewpoint are excluded. Fig. 3 depicts an illustrative example of this algorithm. Another type of algorithm for contour generation is discussed in (Suzuki & Be, 1985).

Die Bundesregierung sagt voraus, daß es 1998 ein Wirtschaftswachstum von bis zu drei Prozent und am Jahresende weniger Arbeitslose geben wird. So steht es im Jahreswirtschaftsbericht, den das Kabinett am Mittwoch verabschiedete. Die Opposition sprach von „Schönfärberei“, Gewerkschaften vom „Prinzip Hoffnung“.



Fig. 2. The bounding rectangles of the page connected components. Image taken from Boiangiu et al. (2008)

Fig. 3: Outcome of the contour detection algorithm. Image taken from Boiangiu et al. (2008)

3.3. Delaunay Triangulation

Following the selection of contours, the constrained Delaunay triangulation algorithm is used to connect all connected components. Nonetheless, this leads to an excessive number of connections. Therefore, a phase of processing is performed to eliminate triangles that connect more than two connected components, leaving only connected components connected in pairs.

As a consequence, this paper establishes by convention two distinct categories of points, namely "current points" and "destination points". These names are derived from their association with or lack of association with a connected component. Using the Delaunay triangulation, multiple triangles emanate from each connected component and extend toward another connected component. Within these triangles, the points located on the current connected component are referred to as current points, while those belonging to triangles located on a different connected component are referred to as destination points.

3.4. Proximity Generation

Proximity is the relationship between two or more connected components. By iteratively processing the triangles present in the constrained Delaunay triangulation and excluding triangles that connect two distinct connected components (inter-triangles), proximity is determined. Excluded from further processing are triangles generated within a single connected component (intra-triangles) or involving three distinct connected components.

The proximity structure furnishes essential details concerning the interrelation between entities, encompassing factors like the paired

connected components, the minimum square distance within Delaunay inter-triangles, the count of connecting points in each connected component, the connection area, and other relevant metrics as dictated by the specific processing needs.

3.5. Separators

As stated in the introduction, separators can be categorized based on their geometrical form or characteristics. Nonetheless, they share a notable characteristic that distinguishes them. Leveraging the aforementioned Delaunay triangulation and the identification of current and destination points enables the application of statistical analysis centered around their ratio.

This examination reveals a fundamental trait of separators: due to their extensive coverage across numerous connected components, regardless of orientation or angle, they exhibit a notably higher count of current points compared to destination points. Thus, separators can be identified, and lines can be drawn between them and conventional characters such as letters and punctuation marks.

The subsequent step involves integrating this data into a hierarchical structure for the page. This integration simplifies the process of recognizing text regions within the hierarchical tree that are enclosed by page boundaries or separators.

4. Cluster Tree

Our methodology involves constructing a hierarchical model using a cluster tree, a specialized form of a multi-way tree. Connected components serve as leaf nodes in this tree structure, while internal nodes represent clusters of connected components. The cluster diameter represents the maximum distance between any two interconnected components within the same cluster, or between neighboring connected components that can form a sequence to link any two such components (see Fig. 4). The principal objective of this tree is to categorize connected components into clusters characterized by expanding diameters. The root of the tree signifies the cluster with the most substantial diameter (if it hypothetically encompassed the entire page, its offspring would correspond to higher-level elements like paragraphs or figures).

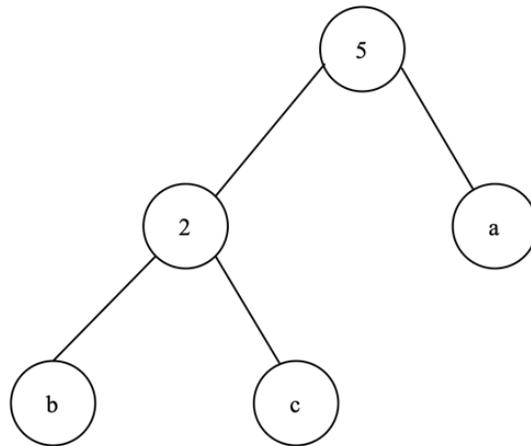


Fig. 4 depicts a simple cluster tree in which the internal nodes are labeled with the diameters of the clusters. Image taken from Boiangiu et al. (2008)

The hierarchical model is combined with the separator data obtained in the preceding stage to ascertain the layout of the page (Gan et al. 2007; Watson, 1981).

There are two possible approaches to designing the cluster tree. The initial approach employs input consisting of extreme points and Delaunay triangulation. Extreme points refer to the points situated along the contours of the connected components.

The algorithm for constructing the tree begins by calculating the minimal length of the Delaunay triangle edge connecting each pair of connected components. The tree is constructed from the ground up. A cluster is created around a random connected component. The connected component closest to the initially connected component is then determined and added to the cluster. Then, the connected component closest to one of the two existing connected components is identified, and if the distance to this connected component is comparable to the distance between the first two, the third connected component is added to the cluster. This procedure is repeated until the nearest connected component has a magnitude so great that it cannot be included in the initial cluster. The remaining clusters are generated using a similar approach, although the algorithm might introduce the nearest cluster instead of the nearest connected component.

The output of the algorithm is the sought-after tree structure that accurately reflects the hierarchy of the page.

„Jobmaschine“

neue Stellen / Opposition: Schönfärberei

ennoch meinte Rexrodt: „liegt hinter uns.“ CDU-Peter Hintze sagte: „Die angesprochen.“

nderungen auf dem Arden laut Rexrodt nur ge- wenn die Koalition ihre Das bedeute mehr Wett- Staat, vereinfachte Pla- und verbesserte Risiko- „Eine wirtschaftspoli- wärts — so wie von der schafft keine Arbeitsplät- die „Fahrt ins rot-grü- würde einem „Katastro- Menschen in Deutsch-

ten Jahreswirtschafts- ment des Starrsinns und igkeit. „Die möglichen Turbulenzen in Südost- tung der Konjunktur in che Zinsanpassungen im mit der Europäischen und die schwache Bin- rden von der Bundesre- rgspeil, anstatt wirt- Vorsorge zu treffen“, kri- illvertretende Fraktions- e Fuchs und der Wirt- rnt Schwanhold.

ditiker bezeichnet die der Regierung als ge-

scheitert. Sie habe zu immer neuen Rek- korden bei der Arbeitslosigkeit, bei der Staatsverschuldung, bei der Steuer- und Abgabenbelastung sowie bei Pleiten ge- führt. „Da nützt auch die Wahlkampfhilfe der Spitzenrepräsentanten aus den Wirt- schaftsv Verbänden nichts“, so die SPD- Abgeordneten.

Auch die Deutsche Angestellten-Ge- werkschaft (DAG) und der Deutsche Ge- werkschaftsbund (DGB) bezweifeln die vorhergesagte Trendwende. Die Arbeitslo- sigkeit werde vielmehr auf Rekordniveau stagnieren, prophezeite die stellvertreten- de DAG-Vorsitzende Ursula Konitzer. Wichtige Wirtschaftsverbände begrüßten dagegen den Bericht. Der Deutsche Indus- trie- und Handelstag nannte ihn „zutref- fend“, der Bundesverband der Deutschen Banken sprach von einer realistischen Darstellung. Zusammen mit anderen Ver- bänden forderten sie aber weitere Refor- men für ein besseres Investitionsklima.

Der Bundesverband mittelständische Wirtschaft (BVMW) kritisierte unterdes- sen die „überleite und weitgehend unkriti- sche“ Parteinahme von Spitzenverbänden der Wirtschaft zugunsten der Regierung. BVMW-Hauptgeschäftsführer Dieter Har- the warnte davor, wie die Gewerkschaften ein politisches Lager offen zu unterstüt- zen. Die Wirtschaftsverbände seien schließlich kein „Kanzlerwahlverein“.

Berichte Seite 4 und Wirtschaftsteil

„Jobmaschine“

neue Stellen / Opposition: Schönfärberei

ennoch meinte Rexrodt: „liegt hinter uns.“ CDU-Peter Hintze sagte: „Die angesprochen.“

nderungen auf dem Arden laut Rexrodt nur ge- wenn die Koalition ihre Das bedeute mehr Wett- Staat, vereinfachte Pla- und verbesserte Risiko- „Eine wirtschaftspoli- wärts — so wie von der schafft keine Arbeitsplät- die „Fahrt ins rot-grü- würde einem „Katastro- Menschen in Deutsch-

ten Jahreswirtschafts- ment des Starrsinns und igkeit. „Die möglichen Turbulenzen in Südost- tung der Konjunktur in che Zinsanpassungen im mit der Europäischen und die schwache Bin- rden von der Bundesre- rgspeil, anstatt wirt- Vorsorge zu treffen“, kri- illvertretende Fraktions- e Fuchs und der Wirt- rnt Schwanhold.

ditiker bezeichnet die der Regierung als ge-

schänert. Sie habe zu immer neuen Re- korden bei der Arbeitslosigkeit, bei der Staatsverschuldung, bei der Steuer- und Abgabenbelastung sowie bei Pleiten ge- führt. „Da nützt auch die Wahlkampfhilfe der Spitzenrepräsentanten aus den Wirt- schaftsv Verbänden nichts“, so die SPD- Abgeordneten.

Auch die Deutsche Angestellten-Ge- werkschaft (DAG) und der Deutsche Ge- werkschaftsbund (DGB) bezweifeln die vorhergesagte Trendwende. Die Arbeitslo- sigkeit werde vielmehr auf Rekordniveau stagnieren, prophezeite die stellvertreten- de DAG-Vorsitzende Ursula Konitzer. Wichtige Wirtschaftsverbände begrüßten dagegen den Bericht. Der Deutsche Indus- trie- und Handelstag nannte ihn „zutref- fend“, der Bundesverband der Deutschen Banken sprach von einer realistischen Darstellung. Zusammen mit anderen Ver- bänden forderten sie aber weitere Refor- men für ein besseres Investitionsklima.

Der Bundesverband mittelständische Wirtschaft (BVMW) kritisierte unterdes- sen die „überleite und weitgehend unkriti- sche“ Parteinahme von Spitzenverbänden der Wirtschaft zugunsten der Regierung. BVMW-Hauptgeschäftsführer Dieter Har- the warnte davor, wie die Gewerkschaften ein politisches Lager offen zu unterstüt- zen. Die Wirtschaftsverbände seien schließlich kein „Kanzlerwahlverein“.

Berichte Seite 4 und Wirtschaftsteil

ie Strafe r nicht bewiesen

medizinischen Attesten 1. Sie sprach während des rlich von „schlechter Be- straft werden müsse.“ nun, daß nach dem Frei- sten die Jugendlichen er- gestellt werden könnten. egen elf der Schüler we- ft in einer terroristischen r ausgesetzt worden, weil e möglicherweise unter gekommen waren. Die ren zu jeweils 15 Jahren orden.

„mittler von amnesty in- than Sugden, forderte in- ktion auf den Freispruch gierungen auf, ihren gierung in Ankara zu er- stärken als zuvor zu Ju- drängung werden. Dies- „verrottet“, sagte Sugden Rundschau. Während des seien die Richter sechs- ltt worden.

ist international bereits Foltner kritisiert worden. gangenen Woche war ein fanisa bekanntgeworden, nach schweren Mißhand- len, seinen vermilten ht zu haben. Drei Tage ter Vater wieder auf. Er 'age bei Verwandten ver-

weiterer Bericht S. 4

Öffentlicher Dienst

Schlichter zeigen sich „gemäßigt optimistisch“

stg BREMEN, 11. März. Die beiden Vorsitzenden der Schlichtungskommission für die schwierige Tarifrunde im Öffentlichen Dienst sind nach eigenen Aus- sagen „gemäßigt optimistisch“, daß sie ein Kompromißpaket schnüren können. Die Tarifparteien seien einigungswillig, sag- ten Bremens Ex-Bürgermeister Hans Kö- schnick (SPD) und der frühere Regie- rungsschef von Rheinland-Pfalz, Carl- Lud- wig Wagner (CDU), am Mittwoch nach dem ersten Tag der Schlichtungsgesprä- che in Bremen. Die Verhandlungen sollen am Montag fortgesetzt werden.

Die „allgemein verfolgte Linie“ läuft nach Angaben der Schlichter darauf hin- aus, die Löhne und Gehälter für die 3,2 Millionen Angestellten und Arbeiter von Bund, Ländern und Gemeinden nur ge- ring zu erhöhen, um Arbeitsplätze zu si- chern. Zudem sollen die Ost-Tarife an West-Niveau angeglichen werden. Kö- schnick und Wagner setzen sich auch da- für ein, steigende Kosten der Alters-Zu- satzversicherung abzufangen. Diese wird bisher nur von den Arbeitgebern finan- ziert. Nach deren Ansicht sollten auch die Beschäftigten Beiträge zahlen.

Die Arbeitgeber haben ein Prozent Lohn- und Gehaltserhöhung angeboten, aber neben dem Zusatzrenten-Beitrag auch Einschnitte bei der Lohnfortzahlung für Kranke und bei Überstundenzuschlä- gen gefordert, was die Gewerkschaften ablehnten. Deren Forderungen würden 4,5 Prozent Mehrausgaben bedeuten.

ie Strafe r nicht bewiesen

medizinischen Attesten 1. Sie sprach während des rlich von „schlechter Be- straft werden müsse.“ nun, daß nach dem Frei- sten die Jugendlichen er- gestellt werden könnten. egen elf der Schüler we- ft in einer terroristischen r ausgesetzt worden, weil e möglicherweise unter gekommen waren. Die ren zu jeweils 15 Jahren orden.

„mittler von amnesty in- than Sugden, forderte in- ktion auf den Freispruch gierungen auf, ihren gierung in Ankara zu er- stärken als zuvor zu Ju- drängung werden. Dies- „verrottet“, sagte Sugden Rundschau. Während des seien die Richter sechs- ltt worden.

ist international bereits Foltner kritisiert worden. gangenen Woche war ein fanisa bekanntgeworden, nach schweren Mißhand- len, seinen vermilten ht zu haben. Drei Tage ter Vater wieder auf. Er 'age bei Verwandten ver-

weiterer Bericht S. 4

Öffentlicher Dienst

Schlichter zeigen sich „gemäßigt optimistisch“

stg BREMEN, 11. März. Die beiden Vorsitzenden der Schlichtungskommission für die schwierige Tarifrunde im Öffentlichen Dienst sind nach eigenen Aus- sagen „gemäßigt optimistisch“, daß sie ein Kompromißpaket schnüren können. Die Tarifparteien seien einigungswillig, sag- ten Bremens Ex-Bürgermeister Hans Kö- schnick (SPD) und der frühere Regie- rungsschef von Rheinland-Pfalz, Carl- Lud- wig Wagner (CDU), am Mittwoch nach dem ersten Tag der Schlichtungsgesprä- che in Bremen. Die Verhandlungen sollen am Montag fortgesetzt werden.

Die „allgemein verfolgte Linie“ läuft nach Angaben der Schlichter darauf hin- aus, die Löhne und Gehälter für die 3,2 Millionen Angestellten und Arbeiter von Bund, Ländern und Gemeinden nur ge- ring zu erhöhen, um Arbeitsplätze zu si- chern. Zudem sollen die Ost-Tarife an West-Niveau angeglichen werden. Kö- schnick und Wagner setzen sich auch da- für ein, steigende Kosten der Alters-Zu- satzversicherung abzufangen. Diese wird bisher nur von den Arbeitgebern finan- ziert. Nach deren Ansicht sollten auch die Beschäftigten Beiträge zahlen.

Die Arbeitgeber haben ein Prozent Lohn- und Gehaltserhöhung angeboten, aber neben dem Zusatzrenten-Beitrag auch Einschnitte bei der Lohnfortzahlung für Kranke und bei Überstundenzuschlä- gen gefordert, was die Gewerkschaften ablehnten. Deren Forderungen würden 4,5 Prozent Mehrausgaben bedeuten.

Fig. 5 The initial image and the outcome of the Delaunay triangulation. Image taken from Boiangiu et al. (2008)

The second approach for constructing the hierarchy involves utilizing coordinates derived from the bounding shapes of the connected components and employing the distance between these bounding shapes as a metric. The convex hull is an appropriate option for the boundary form. In this instance, the convex hull of each connected component is computed using its contour points, and the minimum distance between the bounding structures is determined and used, as before, as the minimum distance between the connected components. Similar to the previous method, the algorithm begins with an empty set and grows a cluster by adding the connected component with the shortest minimum distance between the bounding structures. Clusters are constructed similarly, with the prospect of including other clusters if their distances are of the same magnitude. The principal objective of this tree is to categorize connected components into clusters characterized by expanding diameters. The resulting hierarchical model portrays the page structure, where the root of the tree symbolizes the entire page, and its offspring symbolize top-level layout components like paragraphs, images, tables, titles, headings, and more. The tree's fronds symbolize the smallest elements, which are typically characters.

Fig. 5 illustrates two versions of the identical image: one prior to Delaunay triangulation and the other subsequent to its application in order to provide visual representation and facilitate comprehension. In these images, the connected components are visibly linked by numerous edges, emphasizing their interconnectedness. Each color represents the connections between the points on the boundary of two connected components, providing a distinct indication of their respective distances. As input for the algorithm, only the connection with the shortest distance is selected from this collection.

4.1. Hierarchy Model - Cluster Tree

As mentioned earlier, the page hierarchy can be represented through a cluster tree model. Within this tree structure the terminal nodes symbolize the input-connected components and are organized into clusters. Every cluster is denoted by a tree node with its diameter proportional to its size.

The cluster tree concept assures that the distance between any two elements within a cluster does not exceed the cluster diameter. Hence, each subtree within the arrangement symbolizes a cluster, wherein all nodes share a closer proximity to each other compared to nodes outside the cluster. Fig. 6 depicts an exemplar of a cluster tree arrangement. In this case, the cluster diameters escalate from the bottom to the top of the tree. Each node is a member of exactly one cluster and has no offspring. In the example, the

labels of the non-leaf nodes correspond to the maximal distances within each cluster, denoting the diameter of the cluster. As an illustration, consider the cluster "ab," encompassing the connected components "a" and "b." With a maximum distance of 20, this signifies that no connected components within this cluster are separated by more than 20 units. Moreover, in accordance with the cluster tree definition and inherent in the construction algorithm, no nodes exist within a 21-unit range from either "a" or "b".

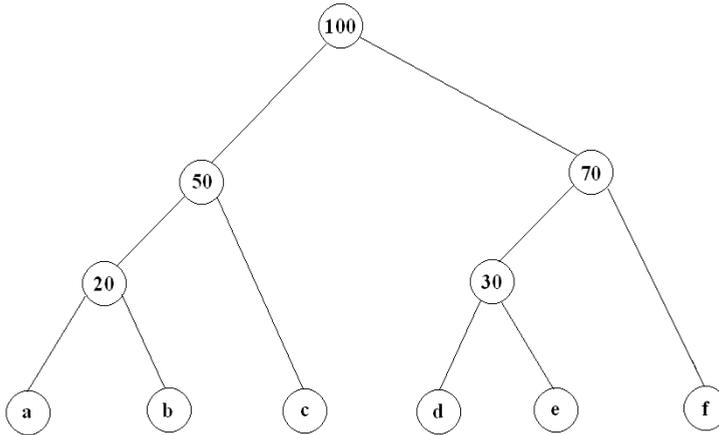


Fig. 6: Illustration of a Cluster Tree.
Image taken from Boiangiu et al. (2008)

If the cluster diameters are not equal, the order of magnitude is considered different in practical implementations. A threshold can be used to determine whether two connected components belong to the same cluster.

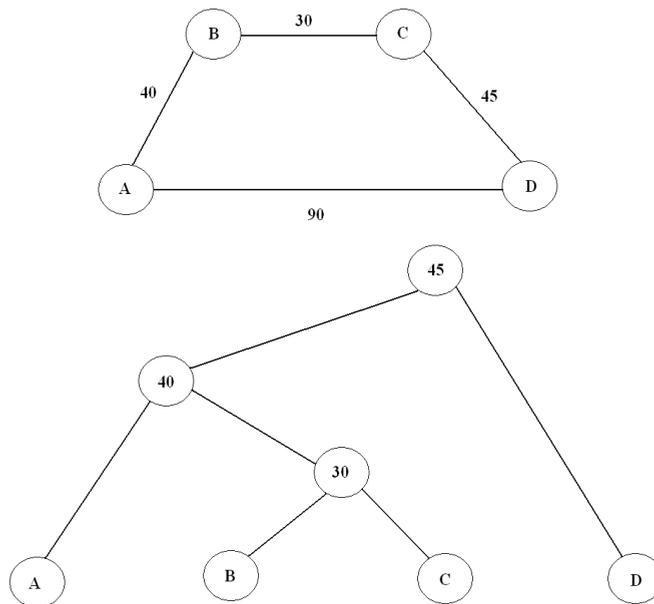


Fig. 7: Illustration of Hierarchy Distance.
Image taken from Boiangiu et al. (2008)

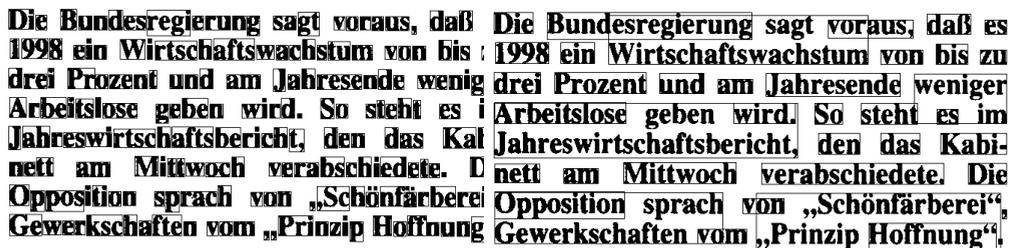
When referring to the distance between two connected components in the context of layout analysis, the diameter of the cluster to which those connected components belong is meant. This measurement is referred to as the hierarchical distance, in contrast to the Euclidean distance utilized for constructing the cluster tree. The Euclidean distance is the minimum path length between two points or the measure of the connecting segment's length. Conversely, the hierarchy distance is understood in a distinct manner. Fig. 7 serves as an illustration of the distinction.

In accordance with the Euclidean distance, the measurement between coordinates "A" and "D" in this illustration is 90 units. The cluster tree discloses that the hierarchy distance from "A" to "D" is 45 units. As "B" and "C" constitute a cluster and subsequently link to "A" in a different cluster before connecting to "D" all three points share an identical hierarchy distance to "D" equivalent to the Euclidean distance between "C" and "D". Consequently, this novel measurement unit is an efficient method for evaluating cluster closure.

Now that the essential terminology has been elucidated, several steps must be taken to obtain the desirable clusters. The initial step involves the formation of the cluster tree. The information contained within the clusters

is subsequently utilized to join distinct connected components or groups of connected components based on hierarchy distances. The following images illustrate the progression of cluster creation and the outcome of dividing the document into zones exhibiting comparable traits.

As observed, this initial iteration only clusters together the nearest connected components. To enhance the clarity of cluster visualization, only a limited section of the original image is employed in the initial set of resultant images, where clusters are denoted by enclosed rectangles. Given the scale of the image, this has a minimal impact and does not aid in zoning the document. In subsequent iterations, however (Figure 8), clusters of connected components are connected, and a hierarchical structure emerges.



Die Bundesregierung sagt voraus, daß 1998 ein Wirtschaftswachstum von bis zu drei Prozent und am Jahresende wenig Arbeitslose geben wird. So steht es im Jahreswirtschaftsbericht, den das Kabinett am Mittwoch verabschiedete. Die Opposition sprach von „Schönfärberei“. Gewerkschaften vom „Prinzip Hoffnung“.

Fig. 8: Presents an initial stage of the algorithm, displaying a notable quantity of clusters stemming from the small minimum value of connected edges at this point. Image taken from Boiangiu et al. (2008)

Fig. 9: The information contained within the clusters is beginning to take shape, forming a semblance of hierarchy. Image taken from Boiangiu et al. (2008)

Through the ongoing process of cluster merging, the target paragraph within the original image is eventually identified as a distinct zone. (Figure 10).

Eventually, each zone is appropriately identified. The process of joining clusters iteratively can proceed, considering the entire page as a cluster on its own. Nonetheless, this might be exorbitant. The objective is to construct zones on the page containing similar information, and the algorithm must terminate after a predetermined number of iterations.

Die Bundesregierung sagt voraus, daß es 1998 ein Wirtschaftswachstum von bis zu drei Prozent und am Jahresende weniger Arbeitslose geben wird. So steht es im Jahreswirtschaftsbericht, den das Kabinett am Mittwoch verabschiedete. Die Opposition sprach von „Schönfärberei“, Gewerkschaften vom „Prinzip Hoffnung“.

Fig. 10: The paragraph has been encompassed within a solitary cluster. Image taken from Boiangiu et al. (2008)

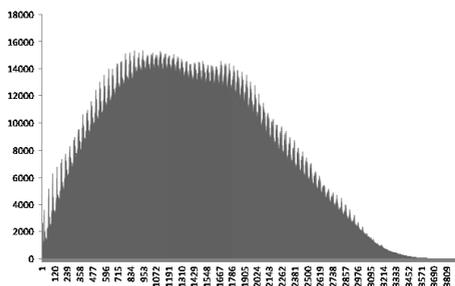


Fig. 11 The histogram of Euclidean distances within the input image. Image taken from Boiangiu et al. (2008)

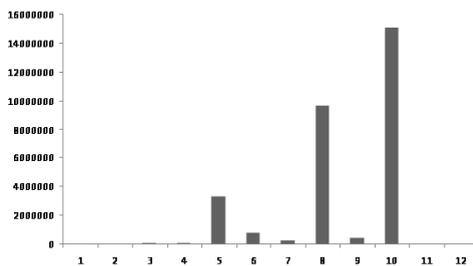


Fig. 12: The histogram of the hierarchical distances computed for the input image. Image taken from Boiangiu et al. (2008)

By plotting histograms of the distance values for the evaluated image, the provided graphics provide an overview of these values. The image's results allow for the detection of titles, paragraphs, and even articles. Nonetheless, without a method for measuring results, determining when to stop iterating becomes uncertain.

Several concepts are introduced in response. Notably, each iteration includes a measure that varies consistently, distinguishing it from all others. Using this information, a system for evaluating results and identifying significant changes can be created.

A critical metric that evolves with each iteration and offers substantial insight into the cluster formation is the rectangular area encompassing the clusters. This metric can be subdivided into three distinct categories: the total rectangular area (computed as the sum of all bounding rectangle areas for the clusters), the overlapping rectangular area (calculated as the sum of all intersecting areas between the bounding rectangles), and the non-overlapping rectangular area (obtained by subtracting the

overlapping area from the total rectangular area). These measurements are illustrated for each iteration in the aforementioned charts.

The analysis of these results allows the identification of inflection points, or points where the slope of the function changes sign. These points are delineated by a white line on the graph, where the function represents the type of area applied to the chart.

Upon careful evaluation of these results, it becomes evident that each inflection point signifies a significant shift in the structure of the chart. For instance, if the subsequent value of the total area exceeds its current value, this indicates a consolidation of clusters into a larger cluster.

CDU beschwört „Jobmaschine“

Bundesregierung **prophezeit** **200.000 neue Stellen** / **Opposition** **Schönfärberei**

Von Rolf-Dietrich Schwarz

Die Bundesregierung sagt voraus, daß es 1998 ein Wirtschaftswachstum von bis zu drei Prozent und am Jahresende weniger Arbeitslose geben wird. So steht es im Jahreswirtschaftsbericht, den das Kabinett am Mittwoch verabschiedete. Die Opposition sprach von „Schönfärberei“ Gewerkschaften vom „Prinzip Hoffnung“.

BONN, 11. März. Die Bundesregierung rechnet in ihrem Jahreswirtschaftsbericht damit, daß bis zum Jahresende 200.000 zusätzliche Arbeitsplätze geschaffen werden. Diese Schätzung liegt weit unter den Versprechen der Wirtschaftsverbände, die 500.000 bis 600.000 neue Stellen angekündigt haben. Bundeskanzler Helmut Kohl (CDU) hatte sich diesem Optimismus angeschlossen. Das Kabinett billigte zudem den Sozialbericht, wonach die Sozialleistungen in den nächsten drei Jahren auf den Stand von Ende der 60er Jahre schrumpfen werden.

Wirtschaftsminister Günter Rexrodt (FDP) prognostizierte, in Westdeutschland werde es zum Jahresende rund 100.000 Arbeitslose weniger geben. Ähnliches könne im Osten erreicht werden. Allerdings bezieht sich die Annahme nur auf die Rechnung von einem Jahresende zum anderen — im Jahreschnitt bleibt die Arbeitslosenquote demnach unverändert bei

11,5 Prozent. Dennoch meinte Rexrodt: „Das Schlimmste liegt hinter uns.“ CDU-Generalsekretär Peter Hintze sagte: „Die Jobmaschine ist angesprungen.“

Die Herausforderungen auf dem Arbeitsmarkt können laut Rexrodt nur gemeistert werden, wenn die Koalition ihre Politik fortsetze. Das bedeute mehr Wettbewerb, weniger Staat, vereinfachte Planungs- und Genehmigungsverfahren, größere Flexibilität und verbesserte Risikokapitalversorgung. „Eine wirtschaftspolitische Rolle rückwärts — so wie von der SPD geplant — schafft keine Arbeitsplätze“, sagte Rexrodt; die „Fahrt ins rote Trau-land“ würde einem „Katastrophen-Trip für die Menschen in Deutschland“ gleichen.

Die SPD nannte den Jahreswirtschaftsbericht ein Dokument des Starrsinns und der Uneinsichtigkeit. „Die möglichen Spätfolgen der Turbulenzen in Südostasien, die Abflachung der Konjunktur in den USA, mögliche Zinsanpassungen im Zusammenhang mit der Europäischen Währungsunion und die schwache Binnennachfrage werden von der Bundesregierung heruntergespielt, anstatt wirtschaftspolitische Vorsorge zu treffen“, kritisierten die stellvertretende Fraktionsvorsitzende Anke Puchs und der Wirtschaftsexperte Ernst Schwanhold.

Beide SPD-Politiker bezeichneten die Angebotspolitik der Regierung als ge-

scheitert. Sie habe zu immer neuen Rekorden bei der Arbeitslosigkeit, bei der Staatsverschuldung, bei der Steuer- und Abgabenbelastung sowie bei Pleiten geführt. „Da nützt auch die Wahlkampfhilfe der Spitzenrepräsentanten aus den Wirtschaftsverbänden nichts“, so die SPD-Abgeordneten.

Auch die Deutsche Angestellten-Gewerkschaft (DAG) und der Deutsche Gewerkschaftsbund (DGB) bezweifeln die vorhergesagte Trendwende. Die Arbeitslosigkeit werde vielmehr auf Rekordniveau stagnieren, prophezeite die stellvertretende DAG-Vorsitzende Ursula Konitzer. Wichtige Wirtschaftsverbände begrüßten dagegen den Bericht. Der Deutsche Industrie- und Handelsstag nannte ihn „zutreffend“, der Bundesverband der Deutschen Banken sprach von einer realistischen Darstellung. Zusammen mit anderen Verbänden forderten sie aber weitere Reformen für ein besseres Investitionsklima.

Der Bundesverband mittelständische Wirtschaft (BVMW) kritisierte indes- den die „uberleite und weitgehend unkritische“ Parteinahme von Spitzenverbänden der Wirtschaft zugunsten der Regierung. BVMW-Hauptgeschäftsführer Dieter Harde warnte davor, wie die Gewerkschaften ein politisches Lager offen zu unterstützen. Die „Wirtschaftsverbände seien schließlich kein „Kanzlerwahlverein“.

Berichte Seite 4 und Wirtschaftsfall

Türkische Polizisten ohne Strafe

Gericht in Manisa nennt Vorwurf der Folter nicht bewiesen

Von Günnar Köhn

ISTANBUL, 11. März. Das Straengericht in der westtürkischen Stadt Manisa hat am Mittwoch zehn Polizisten von dem Vorwurf freigesprochen, sie hätten 16 Jugendliche gefoltert. Dafür gebe es keine hinlänglichen Beweise, hieß es in der Begründung. Einer der Anwälte der Jugendlichen, der sozialdemokratische Abgeordnete Sabri Ergül, sprach von einem „Schock“ und kündigte Berufung gegen das Urteil an. Es müsse sich um einen Justizirrtum handeln.

Die Schüler waren im Dezember 1995 unter dem Verdacht der Mitgliedschaft in einer linksgerichteten militanten Organisation eine Woche lang in der Polizeizentrale von Manisa festgehalten worden. Dort wurden sie nach eigenen Angaben nackt ausgezogen, mit kaltem Wasser abgespritzt und unter anderem an den Genitalien mit Elektroschocks gefoltert. Der damals 14-jährige Mahir Göktaş gab später zu Protokoll: „Sie versetzten mir Stromstöße am rechten Daumen, an meinen Armen, in der Bauchgegend und an meinen Sexualorganen. Danach hatte ich in meinem rechten Fuß und meinen Genitalien überhaupt kein Gefühl mehr.“

Ergül hatte die Jugendlichen bei einem unangemeldeten Besuch der Polizeizentrale zum Teil nackt und mit verbundenen Augen vorgefunden. Doch selbst die Staatsanwaltschaft ließ sich von den Aussagen des Parlamentsmitglieds und von

ten zahlreichen medizinischen Attesten nicht überzeugen. Sie sprach während des Verfahrens lediglich von „schlechter Behandlung“, die bestraft werden müsse.

Ergül fürchtet nun, daß nach dem Freispruch der Polizisten die Jugendlichen erneut vor Gericht gestellt werden könnten. Das Verfahren gegen elf der Schüler wegen Mitgliedschaft in einer terroristischen Organisation war ausgesetzt worden, weil die Geständnisse möglicherweise unter Folter zustande gekommen waren. Die übrigen fünf waren zu jeweils 15 Jahren Haft verurteilt worden.

Der Türkei-Ermittler von amnesty international, Jonathan Sugden, forderte in seiner ersten Reaktion auf den Freispruch ausländische Regierungen auf, ihren Druck auf die Regierung in Ankara zu erhöhen. „Sie muß stärker als zuvor zu Justizreformen gedrängt werden. Dieses System ist völlig verrottet“, sagte Sugden der *Frankfurter Rundschau*. Während des Folterprozesses seien die Richter sechsmal ausgewechselt worden.

Die Türkei ist international bereits mehrfach wegen Folter kritisiert worden. Erst in der vergangenen Woche war ein neuer Fall in Manisa bekanntgeworden. Ein Mann hatte nach schweren Mißhandlungen gestanden, seinen vermißten Vater umgebracht zu haben. Drei Tage später tauchte der Vater wieder auf. Er hatte ein paar Tage bei Verwandten verbracht.

Kommentar S. 3, weiterer Bericht S. 4

Öffentlicher Dienst

Schlichter zeigen sich „gemäßigt optimistisch“

stig BREMEN, 11. März. Die beiden Vorsitzenden der Schlichtungskommission für die schwierige Tarifrunde im Öffentlichen Dienst sind nach eigenen Aussagen „gemäßigt optimistisch“, daß sie ein Kompromißpaket schnüren können. Die Tarifparteien seien einigungswillig, sagte Bremens Ex-Bürgermeister Hans Koschnick (SPD) und der frühere Regierungschef von Rheinland-Pfalz, Carl-Ludwig Wagner (CDU), am Mittwoch nach dem ersten Tag der Schlichtungsgespräche in Bremen. Die Verhandlungen sollen am Montag fortgesetzt werden.

Die „allgemein verfolgte Linie“ läuft nach Angaben der Schlichter darauf hinaus, die Löhne und Gehälter für die 3,2 Millionen Angestellten und Arbeiter von Bund, Ländern und Gemeinden gering zu erhöhen, um Arbeitsplätze zu sichern. Zudem sollen die Ost-Tarife an West-Niveau angeglichen werden. Koschnick und Wagner setzen sich auch dafür ein, steigende Kosten der Alters-Zusatzversorgung abzufangen. Diese wird bisher nur von den Arbeitgebern finanziert. Nach deren Ansicht sollten auch die Beschäftigten Beiträge zahlen.

Die Arbeitgeber haben 10 Prozent Lohn- und Gehaltserhöhung angeboten, aber neben dem Zusatzrentenbeitrag auch Einschnitte bei der Lohnfortzahlung für Kranke und bei Überstundenzuschlägen gefordert, was die Gewerkschaften ablehnen. Deren Forderungen würden 4,5 Prozent Mehrausgaben bedeuten.

Fig. 13 A synopsis of the image at a stage where each presented paragraph is recognized as a distinct cluster. Image taken from Boiangiu et al. (2008)

CDU beschwört „Jobmaschine“

Bundesregierung prophezeit 200 000 neue Stellen / Opposition: Schönfärberei

Von Kolf Dietrich Schwartz
Die Bundesregierung sagt voraus, daß es 1998 ein Wirtschaftswachstum von bis zu drei Prozent und am Jahresende weniger Arbeitslose geben wird. So steht es im Jahreswirtschaftsbericht, den das Kabinett am Mittwoch verabschiedete. Die Opposition sprach von „Schönfärberei“, Gewerkschaften von „Prinzip Hoffnung“.

BONN, 11. März. Die Bundesregierung rechnet in ihrem Jahreswirtschaftsbericht damit, daß bis zum Jahresende 200 000 zusätzliche Arbeitsplätze geschaffen werden. Diese Schätzung liegt weit unter den Versprechen der Wirtschaftsverbände, die 500 000 bis 600 000 neue Stellen angekündigt haben. Bundeskanzler Helmut Kohl (CDU) hatte sich diesem Optimismus angeschlossen. Das Kabinett billigte zudem den Sozialbericht, wonach die Sozialleistungen in den nächsten drei Jahren auf den Stand von Ende der 60er Jahre schrumpfen werden.

Wirtschaftsminister Günter Rexrodt (FDP) prognostizierte, in Westdeutschland werde es zum Jahresende rund 100 000 Arbeitslose weniger geben. Ähnliches könne im Osten erreicht werden. Allerdings bezieht sich die Annahme nur auf die Rechnung von einem Jahresende zum anderen — im Jahreschnitt bleibt die Arbeitslosenquote demnach unverändert bei

11,5 Prozent. Dennoch meinte Rexrodt: „Das Schlimmste liegt hinter uns.“ CDU-Generalsekretär Peter Hintze sagte: „Die Jobmaschine ist angesprungen.“

Die Herausforderungen auf dem Arbeitsmarkt können laut Rexrodt nur gemeistert werden, wenn die Koalition ihre Politik fortsetze. Das bedeute mehr Wettbewerb, weniger Staat, vereinfachte Planungs- und Genehmigungsverfahren, größere Flexibilität und verbesserte Risikokapitalversorgung. „Eine wirtschaftspolitische Rolle rückwärts — so wie von der SPD geplant — schafft keine Arbeitsplätze“, sagte Rexrodt; die „Fahrt ins rot-grüne Traumland“ würde einem „Katastrophen-Trip für die Menschen in Deutschland“ gleichen.

Die SPD nannte den Jahreswirtschaftsbericht ein Dokument des Starrsinns und der Uneinsichtigkeit. „Die möglichen Spätfolgen der Turbulenzen in Südostasien, die Abflachung der Konjunktur in den USA, mögliche Zinsanpassungen im Zusammenhang mit der Europäischen Währungsunion und die schwache Binnennachfrage werden von der Bundesregierung heruntergespielt, anstatt wirtschaftspolitische Vorsorge zu treffen“, kritisierten die stellvertretende Fraktionsvorsitzende Anke Fuchs und der Wirtschaftsexperte Ernst Schwanhold.

Beide SPD-Politiker bezeichneten die Angebotspolitik der Regierung als ge-

scheitert. Sie habe zu immer neuen Rekorden bei der Arbeitslosigkeit, bei der Staatsverschuldung, bei der Steuer- und Abgabenbelastung sowie bei Pleiten geführt. „Da nützt auch die Wahlkampfhilfe der Spitzenrepräsentanten aus den Wirtschaftsverbänden nichts“, so die SPD-Abgeordnete.

Auch die Deutsche Angestellten-Gewerkschaft (DAG) und der Deutsche Gewerkschaftsbund (DGB) bezweifelten die vorhergesagte Trendwende. Die Arbeitslosigkeit werde vielmehr auf Rekordniveau stagnieren, prophezeite die stellvertretende DAG-Vorsitzende Ursula Konitzer. Wichtige Wirtschaftsverbände begrüßten dagegen den Bericht. Der Deutsche Industrie- und Handelstag nannte ihn „zutreffend“, der Bundesverband der Deutschen Banken sprach von einer realistischen Darstellung. Zusammen mit anderen Verbänden forderten sie aber weitere Reformen für ein besseres Investitionsklima.

Der Bundesverband mittelständische Wirtschaft (BVMW) kritisierte unterdessen die „überleite und weitgehend unkritische“ Parteinahme von Spitzenverbänden der Wirtschaft zugunsten der Regierung. BVMW-Hauptgeschäftsführer Dieter Harthe warnte davor, wie die Gewerkschaften ein politisches Lager offen zu unterstützen. Die Wirtschaftsverbände seien schließlich kein „Kanzlerwahlverein“.

Berichte Seite 4 und Wirtschaftsteil

Türkische Polizisten ohne Strafe

Gericht in Manisa nennt Vorwurf der Folter nicht bewiesen

Von Gunnar Köhne

ISTANBUL, 11. März. Das Strafgericht in der westtürkischen Stadt Manisa hat am Mittwoch zehn Polizisten von dem Vorwurf freigesprochen, sie hätten 16 Jugendliche gefoltert. Dafür gebe es keine hinlänglichen Beweise, hieß es in der Begründung. Einer der Anwälte der Jugendlichen, der sozialdemokratische Abgeordnete Sabri Ergül, sprach von einem „Schock“ und kündigte Berufung gegen das Urteil an. Es müsse sich um einen Justizirrtum handeln.

Die Schüler waren im Dezember 1995 unter dem Verdacht der Mitgliedschaft in einer linksgerichteten militanten Organisation eine Woche lang in der Polizeizentrale von Manisa festgehalten worden. Dort wurden sie nach eigenen Angaben nackt ausgezogen, mit kaltem Wasser abgespritzt und unter anderem an den Genitalien mit Elektroschocks gefoltert. Der damals 14-jährige Mahr Göktaş gab später zu Protokoll: „Sie versetzten mir Stromstöße am rechten Daumen, an meinen Sexualorganen. Danach hatte ich in meinem rechten Fuß und meinen Genitalien überhaupt kein Gefühl mehr.“

Ergül hatte die Jugendlichen bei einem unangemeldeten Besuch der Polizeizentrale zum Teil nackt und mit verbundenen Augen vorgefunden. Doch selbst die Staatsanwaltschaft ließ sich von den Aussagen des Parlamentsmitglieds und von

den zahlreichen medizinischen Attesten nicht überzeugen. Sie sprach während des Verfahrens lediglich von „schlechter Behandlung“, die bestraft werden müsse.

Ergül fürchtet nun, daß nach dem Freispruch der Polizisten die Jugendlichen erneut vor Gericht gestellt werden könnten. Das Verfahren gegen elf der Schüler wegen Mitgliedschaft in einer terroristischen Organisation war ausgesetzt worden, weil die Geständnisse möglicherweise unter Folter zustande gekommen waren. Die übrigen fünf waren zu jeweils 15 Jahren Haft verurteilt worden.

Der Türkei-Ermittler von amnesty international, Jonathan Sugden, forderte in einer ersten Reaktion auf den Freispruch ausländische Regierungen auf, ihren Druck auf die Regierung in Ankara zu erhöhen. „Sie muß stärker als zuvor zu Justizreformen gedrängt werden. Dieses System ist völlig verrottet“, sagte Sugden der *Frankfurter Rundschau*. Während des Folterprozesses seien die Richter sechsmal ausgewechselt worden.

Die Türkei ist international bereits mehrfach wegen Folter kritisiert worden. Erst in der vergangenen Woche war ein neuer Fall in Manisa bekanntgeworden. Ein Mann hatte nach schweren Mißhandlungen gestanden, seinen vermißten Vater umgebracht zu haben. Drei Tage später tauchte der Vater wieder auf. Er hatte ein paar Tage bei Verwandten verbracht.

Kommentar S. 3, weiterer Bericht S. 4

Öffentlicher Dienst

Schlichter zeigen sich „gemäßigt optimistisch“

11. März. Die beiden Vorsitzenden der Schlichtungskommission für die schwierige Tarifrunde im Öffentlichen Dienst sind nach eigenen Aussagen „gemäßigt optimistisch“, daß sie ein Kompromißpaket schnüren können. Die Tarifparteien seien einigungswillig, sagten Bremens Ex-Bürgermeister Hans Koschnick (SPD) und der frühere Regiererschef von Rheinland-Pfalz, Carl-Ludwig Wagner (CDU), am Mittwoch nach dem ersten Tag der Schlichtungsgespräche in Bremen. Die Verhandlungen sollen am Montag fortgesetzt werden.

Die „allgemein verfolgte Linie“ läßt nach Angaben der Schlichter darauf hinaus, die Löhne und Gehälter für die 3,2 Millionen Angestellten und Arbeiter von Bund, Ländern und Gemeinden nur gering zu erhöhen, um Arbeitsplätze zu sichern. Zudem sollen die Ost-Tarife an West-Niveau angeglichen werden. Koschnick und Wagner setzen sich auch dafür ein, steigende Kosten der Alters-Zustatzversorgung abzufangen. Diese wird bisher nur von den Arbeitgebern finanziert. Nach deren Ansicht sollten auch die Beschäftigten Beiträge zahlen.

Die Arbeitgeber haben ein Prozent Lohn- und Gehaltssteigerung angeboten, aber neben dem Zusatzrenten-Beitrag auch Einschnitte bei der Lohnfortzahlung für Kranke und bei Überstundenzuschlägen gefordert, was die Gewerkschaften ablehnen. Deren Forderungen wurden 4,5 Prozent Mehrausgaben bedeuten.

Fig. 14: The final outline, where all significant regions of the page are each identified within a distinct cluster.

Image taken from Boiangiu et al. (2008)

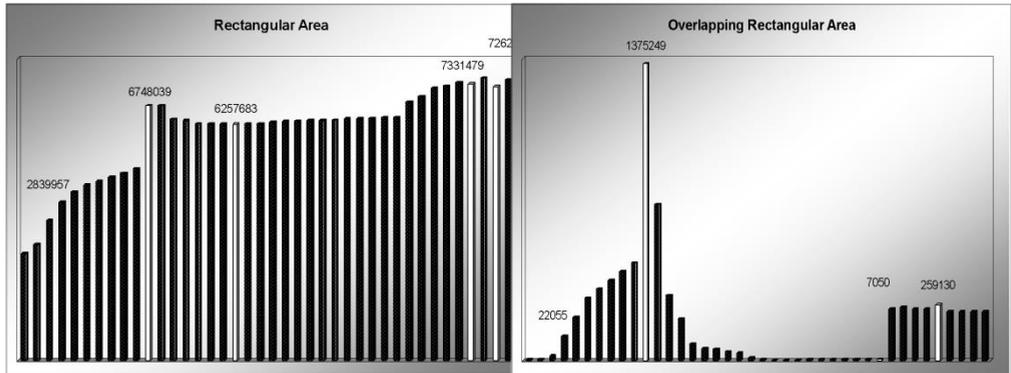


Fig. 15: This chart represents the changes in rectangular area values for each iteration performed on the tested image. Image taken from Boiangiu et al. (2008)

Fig. 16: This chart illustrates the variations in overlapping rectangular area values across all iterations performed on the tested image. Image taken from Boiangiu et al. (2008)

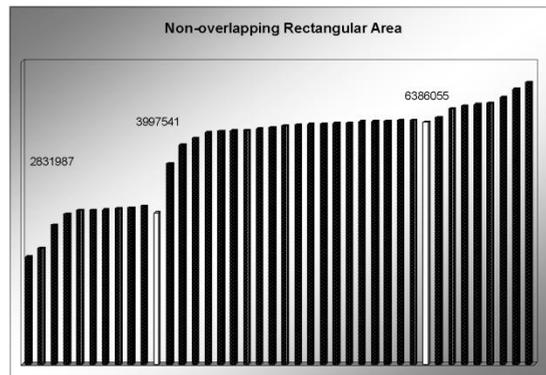


Fig. 17: This chart depicts the changes in non-overlapping rectangular area values throughout all iterations conducted on the tested image. Image taken from Boiangiu et al. (2008)

Conversely, a decrease in the subsequent value compared to the current one signifies that certain clusters which were initially separate have amalgamated into a larger cluster, leading to a reduction in the overall area. By tracking the shifts in slope sign from positive to negative and vice versa, it becomes evident that the most pronounced changes occur precisely at these junctures. Therefore, the decision to terminate the algorithm at a

specific iteration should be informed by these points. For the attainment of the most optimal cluster hierarchy, it is advisable to designate one of the terminal inflection points as the termination point for the algorithm.

5. Metrics for determining where to extract from a cluster tree the layout elements

When determining where to extract layout elements from a cluster tree in document image analysis, various metrics can be considered.

One such metric is the spatial arrangement of pixels. In the context of a document image, spatial arrangements can help identify the structural relationships between different elements, Kise et al. (1998). Titles generally inhabit the uppermost region of a page, while the body, constituting paragraphs and images, spans the top to the bottom (Haralick & Shapiro, 1992). Analyzing spatial relationships between clusters can provide insights into potential layout elements.

Another essential metric is cluster density. Dense clusters frequently correspond to text elements like paragraphs and titles, characterized by high-density pixel collections (Doermann & Tombre, 1998). In contrast, lower-density clusters might signify image elements or spaces between different layout components (Boiangiu et al., 2008). The cluster's density can be quantified using several statistical measures, such as variance or standard deviation.

The size and shape of clusters are also influential in determining the layout elements. Larger clusters may indicate extensive bodies of text, like paragraphs, while reduced clusters could point toward titles or bullet points (Louloudis et al. 2009). The cluster's shape, determined by the pixel arrangement within a cluster, can help distinguish between linear text elements and more intricate graphic elements (Chen et al., 1999).

A pivotal concept in cluster trees is the cluster level. Higher-level clusters usually represent larger, general layout elements, while lower-level clusters correspond to more specific elements (Sarkar et al., 2004). Examining the level of a cluster in the tree can provide insights into the granularity of the layout element that the cluster represents.

The interconnectivity of clusters within a cluster tree can offer additional insights. Clusters closely interconnected are likely part of the same layout element or closely related elements (Kong et al., 2010). On the other hand, clusters with fewer or weaker connections might represent discrete elements (Kong et al., 2010).

These metrics are not used isolated but are combined and weighted appropriately to decide where to extract layout elements from a cluster tree,

(Antonacopoulos, 1998). The best combination of these metrics is contingent on the specific document image and the goals of the analysis.

As a result, the rectangular area taken up by the clusters becomes the pivotal variable that changes with each iteration and offers the most informative insights. This can be separated into three categories: total rectangular area, overlapping rectangular area, and rectangular area without overlap. These measurements are depicted in the aforementioned charts for each iteration.

By evaluating the results, we can identify inflection points on the graph, or points where the function's slope changes signs. In this instance, the function corresponds to the chart area category. These inflection points, denoted on the graph by white lines, indicate significant changes. For instance, if the subsequent total area value surpasses the current value, it signifies the merging of clusters into a larger entity.

If the subsequent value is smaller than the current one, it indicates the integration of certain clusters that were previously part of a larger cluster, leading to a decrease in the overall area. Observing the changes in slope sign from ascending to descending and vice versa, we find that the most substantial changes occur precisely at these points. Therefore, the decision to terminate the algorithm must take into account these inflection points, with one of the final inflection points serving as the termination point in order to acquire the optimal cluster hierarchy.

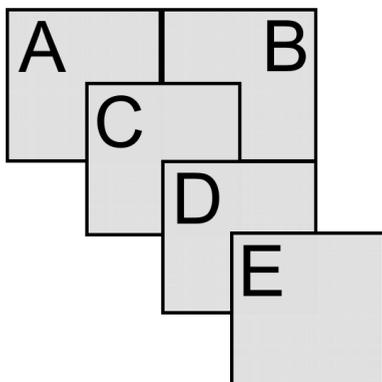


Fig. 18: This figure provides an illustration of the rectangular area measurement application.
Image taken from Boiangiu et al. (2008)

Figure 18 illustrates the use of rectangular area measurement with an example. In this case, the total rectangular area is 20 units, the overlapping rectangular area is 4 units, and the non-overlapping rectangular area is 11

units (calculated by subtracting the overlapping area from the sum of rectangle areas, presuming each rectangle has a 2-by-2-unit area).

This study presents an effective instrument for page layout analysis that permits the selection of distinct groups of connected components by slicing the tree at various levels. Through this method, paragraphs, headings, and other layout elements can be precisely detected using a simple and easily implementable algorithm. In addition, the method has prospective applications outside of document content conversion.

Through the use of mathematical solutions, algorithms, and content analysis, the efficacy of this strategy can be readily demonstrated and confirmed. The method of layout analysis described in this paper is a natural progression of hierarchical clustering. It simulates the progressive withdrawal of the viewer from a document, resulting in an image that becomes increasingly "blurrier." Despite the loss of specifics, the overall structure of the document is still discernible, including the positioning of paragraphs, headings, tables, and images.

Employing Delaunay triangulation structures ensures the preservation of precision throughout multiple levels of resampling, thereby faithfully capturing the phenomenon of grouping (clustering) perceived by the human eye as viewing distance expands.

Moreover, the human visual system exhibits greater sensitivity to structures characterized by a rectangular shape. The prioritization of rectangular reconstruction of clusters within the document is achieved through the utilization of cluster-area functions that emphasize local maximization.

By integrating multiple clustering measures, the algorithm is able to determine the optimal strategy for a given document layout.

6. Postprocessing Phase of Layout Analysis via Non-Maximum Suppression

Once the document image analysis has generated a preliminary layout, a vital postprocessing stage is required to refine the result, effectively removing potential overlapping elements and ensuring the best representation of layout components. This process utilizes Non-Maximum Suppression (NMS), a technique prevalently employed in computer vision tasks to reduce the number of overlapping bounding boxes and retain the most suitable ones (Malisiewicz et al., 2011; Ren et al. 2017).

In this study, three variants of NMS were examined: the classic approach, linear NMS, and Gaussian NMS. Each of these variants has

unique characteristics and applications, necessitating a comprehensive evaluation to determine the most effective for this specific task.

The classic NMS, often applied in object detection tasks, operates by retaining the bounding box with the highest confidence score while suppressing all others that significantly overlap with it, Ren et al. (2017). However, its rigid suppression rule may lead to the omission of valid boxes in close proximity, affecting the representation of closely packed document elements (Bodla et al., 2017).

Linear NMS, introduced by Hosang et al. (2016), offers a more flexible approach by gradually decreasing the scores of suppressed boxes based on their overlap with the retained box. However, it still relies on a deterministic suppression function, which may not always be suitable for complex document layouts.

In contrast, Gaussian NMS, proposed by He et al. (2019), employs a probabilistic approach by reducing the scores of suppressed boxes using a Gaussian function, offering a gentler and more nuanced suppression.

The performance of these NMS variants was evaluated on a custom dataset consisting of 121 documents from various epochs, with diverse layouts. This approach was intended to capture a wide range of layout complexities and variations, mirroring real-world scenarios (Sharma et al. 2013).

Following a comprehensive evaluation, Gaussian NMS emerged as the most effective variant for the task, given the fact that it ensured a more robust thresholding range for filtering out unwanted blocks. Despite the complexity and diversity of the dataset, Gaussian NMS consistently provided superior results. Its probabilistic and gentle suppression approach proved particularly advantageous in handling various layout elements that exist in close proximity, a common occurrence in complex documents. Consequently, Gaussian NMS emerged as the optimal choice for the postprocessing phase of document image layout analysis in our work.

7. Conclusion

The methodology delineated herein provides an effective instrument for analyzing page layouts, facilitating the segregation of different entity groups by trimming the cluster tree at varying levels. This approach enables the precise detection of paragraphs, headings, and other layout elements via a straightforward, easily implementable algorithm. Furthermore, the utility of this method extends beyond the realm of document content conversion.

The validity of the proposed approach can be readily substantiated and verified by employing a combination of mathematical solutions, algorithms, and conventional content analysis techniques.

The layout analysis approach outlined in this study represents a logical advancement of hierarchical clustering procedures. Envision the scenario in which an observer progressively withdraws from a document. The resultant image becomes increasingly nebulous, and while the fine details become indistinguishable, the document's overall structure, including the placement of paragraphs, headings, tables, and images, remains discernible.

With an increase in distance from the document, the observer discerns fewer intricacies of its content while gaining a broader perspective of its overall layout. This process can be replicated by employing a pyramid-like resampling of the image, gradually reducing it until it eventually contracts into a single point. This intuitive procedure is mathematically formalized through Delaunay triangulation structures, guaranteeing the preservation of precision across different resampling stages. Consequently, the clustering of elements mirrors the human eye's behavior when increasing the viewing distance from the document.

Additionally, the human eye exhibits heightened sensitivity to rectangular-like structures. Hence, the emphasis is placed on the rectangular reconstruction of clusters within the document through the utilization of specific cluster-area functions that require local maximization.

Offering a range of clustering measures empowers the clustering algorithm to opt for the most appropriate approach based on the specific layout of the document.

References

- Antonacopoulos, A. (1998). Page segmentation using the description of the background. *Computer Vision and Image Understanding: CVIU*, 70(3), 350–369. doi:10.1006/cviu.1998.0691
- Baird, H. S. (2002). Document image defect models and their uses. *Proceedings of 2nd International Conference on Document Analysis and Recognition (ICDAR '93)*. Tsukuba Science City. doi:10.1109/icdar.1993.395781
- Baird, H. S. (2005). The State of the Art of Document Image Degradation Modeling,” in *Document Analysis Systems VI DAS 2004*. Lecture Notes in Computer Science, B. H. and S. A.L, Eds., Berlin, Heidelberg: Springer.
- Bodla, N., Singh, B., Chellappa, R. & Davis, L. S. (2017). Soft-NMS -- improving object detection with one line of code. <http://arxiv.org/abs/1704.04503>

- Boiangiu, C. A., Cananau, D. C. & Bucur, I. (2008). A Hierarchical Clustering Method Aimed at Document Layout Understanding and Analysis. *International Journal of Mathematical Models and Methods in Applied Sciences*, 2(1), 413–422.
- Chen, H., Bloomberg, D. S. & Baird, H. S. (1999). *Document Image Defect Models and Their Uses*, in *Document Analysis Systems. DAS 1998. Lecture Notes in Computer Science*. Springer.
- Coustaty, M., Bertet, K., Visani, M. & Ogier, J. M. (2011). A New Multi-layered Bitmap Model for Document Image Analysis. *2011 International Conference on Document Analysis and Recognition*.
- Doermann, D. & Tombre, K. (1998). *Progress in Pattern Recognition, Image Analysis and Applications*. Springer-Verlag.
- Gan, G., Ma, C. & Wu, J. (2007). *Data Clustering: Theory, Algorithms, and Applications*. SIAM.
- Gong, Y. & Liu, X. (2016). Document Clustering via Matrix Representation. *2016 International Conference on Pattern Recognition (ICPR)*.
- Guibas, L. J. & Stolfi, J. (1985). Primitives for the manipulation of general subdivisions and the computation of Voronoi diagrams. *ACM Transactions on Graphics*, 11(3). doi:10.1145/800061.808751
- Haralick, R. M. & Shapiro, L. G. (1992). *Computer and Robot Vision, Volume II*. Upper Saddle River, NJ: Pearson.
- He, P., Cai, Z., Tian, X. & Zuo, W. (2019). AP Loss for Multi-box Detection. In *Proceedings of the British Machine Vision Conference*. BMVC.
- Hosang, J., Benenson, R., Dollár, P. & Schiele, B. (2016). What makes for effective detection proposals? *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(4), 814–830. doi:10.1109/TPAMI.2015.2465908
- Jain, A. K. & Yu, B. (1998). Document representation and its application to page decomposition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(3), 294–308. doi:10.1109/34.667886
- Jain, A.K. & Dubes, R.C. (1988). *Algorithms for clustering data*. Prentice-Hall
- Kise, K., Sato, A. & Iwata, M. (1998). Segmentation of page images using the area Voronoi diagram. *Computer Vision and Image Understanding: CVIU*, 70(3), 370–382. doi:10.1006/cviu.1998.0684
- Kong, Y., Franke, K. & Rosenhahn, B. (2010). Using Graph Cuts for Document Layout Extraction. *Advances in Visual Computing. ISVC 2010*. Springer.
- Lee, D. T. & Lin, A. K. (1986). Generalized delaunay triangulation for planar graphs. *Discrete & Computational Geometry*, 1(3), 201–217. doi:10.1007/bf02187695

- Lee, D. T. & Schachter, B. J. (1980). Two algorithms for constructing a Delaunay triangulation. *International Journal of Computer & Information Sciences*, 9(3), 219–242. doi:10.1007/bf00977785
- Louloudis, G., Gatos, B., Pratikakis, I. & Halatsis, C. (2009). Text line and word segmentation of handwritten documents. *Pattern Recognition*, 42(12), 3169–3183. doi:10.1016/j.patcog.2008.12.016
- Malisiewicz, T., Gupta, A. & Efros, A. A. (2011). Ensemble of exemplar-SVMs for object detection and beyond. *2011 International Conference on Computer Vision*. doi:10.1109/iccv.2011.6126229
- Nagy, G., Seth, S. & Viswanathan, K. (2004). A prototype document image analysis system for technical journals. *Computer*, 7(25), 10–22.
- Otsu N. (1979). A threshold selection method from gray-level histograms. *IEEE Transactions on Systems, Man, and Cybernetics*, 9(1), 62–66.
- Qumsiyeh, R. J. (1995). Line detection in document images. In *International Conference on Image Processing*. 477–480.
- Ren, S., He, K., Girshick, R. & Sun, J. (2017). Faster R-CNN: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(6), 1137–1149. doi:10.1109/TPAMI.2016.2577031
- Sarkar, P. & Bhowmick, P. (2011). An Approach to Document Image Block Classification and Grouping Using ICA. *2011 International Conference on Document Analysis and Recognition*.
- Sarkar, Prateek, Baird, H. S. & Zhang, X. (2004). Training on severely degraded text-line images. *Seventh International Conference on Document Analysis and Recognition, 2003. Proceedings*. doi:10.1109/icdar.2003.1227624
- Sauvola, J. & Pietikäinen, M. (2000). Adaptive document image binarization. *Pattern Recognition*, 33(2), 225–236 doi:10.1016/s0031-3203(99)00055-2.
- Sharma, A., Ojha, U. & Govindaraju, V. (2013). Adapting state-of-the-art printed document layout analysis techniques for handwriting identification. *Pattern Recognition*, 46(3), 881–895.
- Suzuki, S. & Be, K. (1985). Topological structural analysis of digitized binary images by border following. *Computer Vision, Graphics, and Image Processing*, 30(1), 32–46. doi:10.1016/0734-189x(85)90016-7
- Wang, J., Bourbakis, N. G. & Triantafyllidis, G. A. (1999). Document image analysis using Voronoi tessellation. *IEEE International Conference on Systems, Man, and Cybernetics*, 1, 230–234.
- Watson, D. F. (1981). Computing the n-dimensional Delaunay tessellation with application to Voronoi polytopes. *The Computer Journal*, 24(2), 167–172. doi:10.1093/comjnl/24.2.167

- Wenyin, L. & Dori, D. (1997). A protocol for performance evaluation of line detection algorithms. *Machine Vision and Applications*, 9(5–6), 240–250. doi:10.1007/s001380050045
- Zhu, G. & Doermann, D. (2007). Automatic Document Logo Detection. *Ninth International Conference on Document Analysis and Recognition (ICDAR 2007)*, 2. doi:10.1109/icdar.2007.4377038