Marius Bulacu, Rutger van Koert, Lambert Schomaker, Tijn van der Zant (2007) "Layout analysis of handwritten historical documents for searching the archive of the Cabinet of the Dutch Queen", Proc. of 9th Int. Conf. on Document Analysis and Recognition (ICDAR 2007), IEEE Computer Society, pp. 357-361, vol. I, 23 - 26 September, Curitiba, Brazil.

Layout Analysis of Handwritten Historical Documents for Searching the Archive of the Cabinet of the Dutch Queen

Marius Bulacu Rutger van Koert Lambert Schomaker Tijn van der Zant
Artificial Intelligence Institute, University of Groningen, The Netherlands
{bulacu, R.van.Koert, schomaker, tijn}@ai.rug.nl

Abstract

In this paper, we describe the structure and the performance of a layout analysis system developed for processing the handwritten documents contained in a large historical collection of very high importance in the Netherlands. We introduce a method based on contour tracing that generates curvilinear separation paths between text lines in order to preserve the ascenders and descenders. Our methods are relevant to research on digitization and retrieval of handwritten historical documents.

1. Introduction

In recent years, the document analysis and recognition community has shown increasing interest in the processing of historical documents. These old documents often have historical and cultural significance and the aim is to scan them and create digital libraries [2], thereby offering continuous electronic access to this important part of the cultural heritage. The challenge is to create automatic search engines that allow the users to find and retrieve only the documents with the relevant content from the entire collection. After this content-based selection is performed, the users are offered access to the scanned images to be able to fully research the original documents and also directly appreciate their graphical beauty. Our long-term goal is to build such an automatic search engine for a historical collection of handwritten documents of high importance in the Netherlands.

In this paper, we describe the layout analyzer that we built as the first step towards the automatic content-based retrieval of document images in our historical collection. Layout analysis for handwritten documents [4, 6, 12] is more difficult and must be conducted in a different way than for machine-print documents [3, 5, 8, 10]. Notably, connected components, the work-horse of machine-print recognition [7], will perform poorly on our handwritten documents because ascenders and descenders touch across several text lines. This observation together with the specific characteristics of our documents (as we will describe further) have lead us to chose projection profiles as the main analysis method combined with a number of other image processing techniques (color filtering, contour tracing and run-length extraction).

The current paper represents a study-case in layout analysis of handwritten documents. We also construct a simple and robust line separation method (the “droplet” technique) that crystallizes ideas that have been present in the research community, but have not yet found a clear reference point until the present.

2. The KdK Collection

The historical document collection at the core of our research effort is the archive of the cabinet of the Dutch Queen – Kabinet der Koningin (KdK) [1]

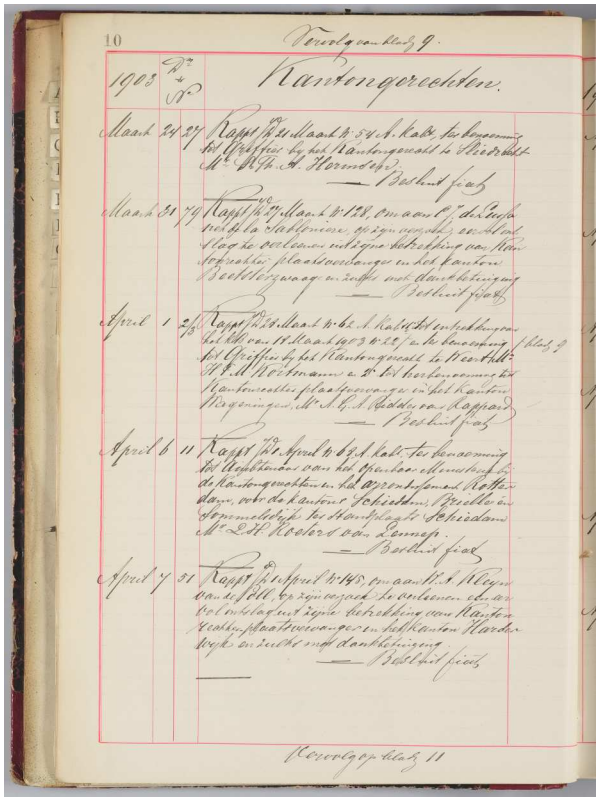

Figure 1: A typical full page from the KdK index.

The KdK collection spans the years 1798 – 1988 and it is in the custody of the *Nationaal Archief* in the Netherlands. This collection is very important historically because it holds all the correspondence between the members of the government and the Queen: all important forms of government intervention (such as laws, decisions, official appointments) need to be formally approved by the Queen and are dutifully collected and recorded in the KdK archive. The official documents treat topics such as: matters of state finance, appointments of judges, administration of the army, state pensions, administration of the colonies, diplomatic relations, taxes etc. People from all over the Netherlands consult these documents for historical or genealogical studies. The KdK collection represents a central entry-point to further searching all other archives of the Dutch government.

An important aspect of this collection is that an index has been carefully kept throughout the two centuries over which the archive was accumulated. The subject of our study is precisely this index of the KdK collection. The index is handwritten in a fixed-format table and, remarkably, the document layout has, for the most part, remained the same over the 200 year period of the collection. Figure 1 shows the image of a typical page from the index.

Every government decision deposited in the large KdK archive has a corresponding entry in the index: the date and number of the decision are written in the first three columns of the table, followed, in the fourth column (wide), by a paragraph describing in short the content of the decision, while the fifth column is occasionally used for special observations. In the table header (see Fig. 1) are recorded: the year (1903), a constant graphical element (a paraph) and the general topic covering all decisions recorded on that particular index page (in the example, “Kantongerechten” – translates to “local courts of law”). The page header contains a machine imprint with the page number and a handwritten reference to the previous index page, while the footer points to the next index page. The complete index of the KdK collection comprises approximately 300 thousand handwritten pages.

The *aim of our layout analysis* is to segment the image into rectangular boxes corresponding to all the document elements described here. There are a number of *important properties of our documents* that have influenced the choice of image processing techniques in building our layout analyzer:

- given their administrative nature, the documents are very orderly and their structure was rigidly maintained such that a top-down model can be effectively imposed to guide the analysis process and also to reject many of the incorrect layout solutions, thus improving the performance of the system;
- the documents have almost no skew and the handwritten lines are always nearly horizontal and have equal spacing, therefore horizontal projection profiles can work well;
- in the KdK collection, the table in which the index entries are written is drawn in a very saturated red color that can be effectively used to find the boundaries between the table columns and cells and therefore split the document image into semantically meaningful regions of interest;
- the writer uses very tall ascenders and very deep descenders that would be damaged by a straight separator between successive handwritten lines, therefore it becomes necessary to use curvilinear separation paths between text lines.

It is important to observe here that the decision paragraphs are the essential object of interest for the users who wish to retrieve, from the KdK collection, those paragraphs containing specified keywords. On the other hand, for the developers of the system, the handwritten text line represents the essential object of interest because it is the input to the subsequent pattern recognition process.

3. Layout Analysis of the KdK collection

Layout analysis is performed in two main stages. The first stage generates the course layout of the document by finding the page borders, the rule lines of the index table and the handwritten text lines grouped into decision paragraphs. Until this point, straight separators are used between the text lines. The correctness of these results is visually checked. In the second stage of the analysis we run the “droplet” method: the straight separators are used as guides for generating curvilinear separation boundaries using contour tracing to go around the ascenders and descenders (when this is possible), in order to avoid damaging the ink.

3.1 Course Layout

The course layout of our documents is generated by the following sequence of techniques.

- **Finding the rule lines of the table and the page margins.**

The most salient element of our documents that can be most robustly found is the index table. We tested two methods: the first uses color information, while the second is more in line with traditional document processing and takes as input gray-scale images.

In the first method, a binary image is generated in which the pixels are set to "on" when they have a highly saturated red color (the red channel exceeds the green and blue ones by a predefined threshold). The positions of the rule lines are then found by computing the horizontal and vertical projections of the binary image.

In the second method, the gray-scale images are binarized using Otsu's method [9] and the vertical rule lines of the table are found by detecting long run-lengths of ink. This is equivalent to the convolution with a matched line detector. The horizontal lines are then found by extracting run-lengths in the image region corresponding to the different columns of the table.

The page margins are located by detecting large variations of the gray levels at a predefined range of distances outside the index table, under the assumption that the documents have been scanned on a dark background.

- **Finding the handwritten text lines and the decision paragraphs.**

The handwritten text lines contained in the main column of the table are found by detection peaks and valleys in the smoothed horizontal projection profile of the binarized image. Only the image region corresponding to the central column is used.

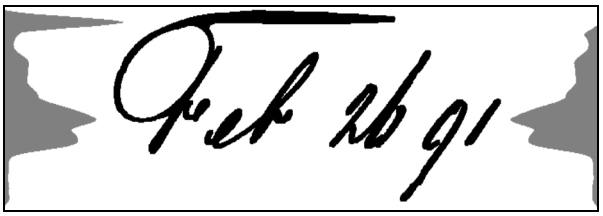

Figure 2: On the left is a standard projection profile, while on the right the ink run-length was used for normalization. Notice that the sharp top peak corresponding to the horizontal dash is eliminated.

We employ a special technique to make the projections reflect the number of ink-paper transitions, rather than the total amount of ink accumulated in a horizontal run. While computing the projection, the algorithm keeps count of the ink run-length, which is then used to normalize the ink sum. By using this technique, we avoid (over)detecting, as separate lines, the sharp peaks due to the occasional handwritten horizontal dashes (see Fig. 2).

The presence of ink in the first three columns of the table marks the end of the old paragraph and the beginning of a new one. The vertical positions of these paragraph headings are then used to group the handwritten lines of the central column into decision paragraphs (see Fig.1 and Fig. 4).

- **Applying top-down checks on the spacing between the found elements.**

Because our documents have a rather rigid format, checking the distances between the found layout elements allows automatically discovering the doubtful situations and selecting them for manual inspection.

3.2 Fine layout – the droplet line segmentation method

In our collection, using straight cuts to separate the text lines would damage nearly all ascenders and descenders, creating many loose components difficult to accommodate in the subsequent pattern recognition process. We therefore use the "droplet" method to obtain curvilinear line separators that preserve most of the ink connectivity.

An intuitive description of this line-finding algorithm that also explains its name would be the following (see Fig. 3). Consider the document turned by 90 degrees such that the text lines run vertically. Then imagine a water droplet falling from the top border of the image starting at an initial position centered between two text lines. The droplet goes straight down unless it hits ink. When this happens, the droplet tries to go around the ink, left or right, often also going against gravitation. When both directions fail, the droplet dissolves the ink along the initial straight path and continues its fall. The algorithm stops when the droplet reaches the bottom of the image.

Where ink cuts have to be applied, the situation is marked such that this potentially useful information can be made available to a recognizer at a later stage.

Figure 3: The "droplet" technique generates curvilinear line separators that preserve the ascenders and descenders. Occasionally, when both search directions fail, ink cuts are applied.

Moore's algorithm is used to follow the contour of the ink. Clockwise and anticlockwise radial sweeps are used in order to implement the two search directions. A failure is declared when the search path goes beyond the location of the maximum of the projection profile (roughly corresponding to the mid position of a text line). A limit at an adjustable distance from the projection maximum (or other heuristics) can be used, depending on the nature of the analyzed documents.

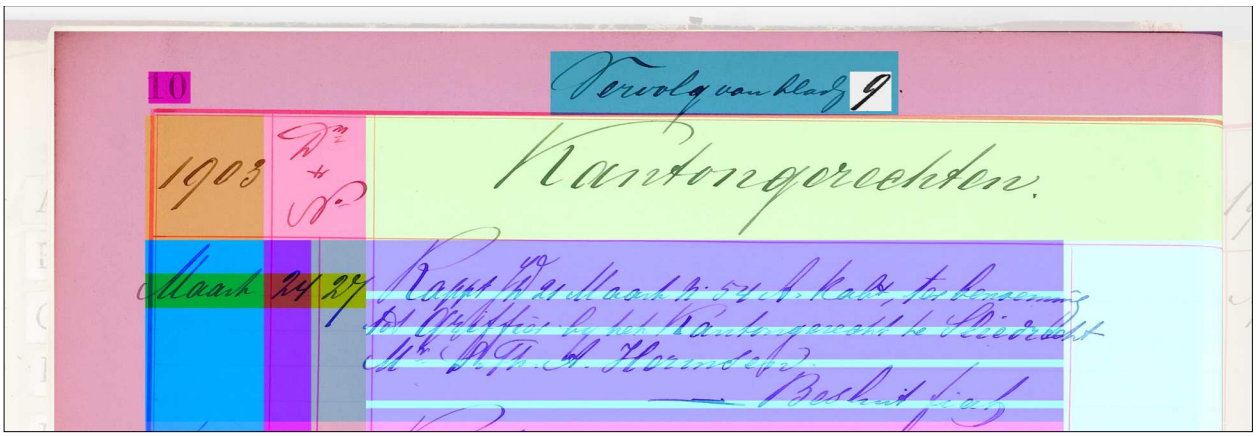

Figure 4: Example of layout analysis results. A colorization scheme was used to mark the different regions.

The ink is dilated prior to using this method in order to keep the droplet path away from the written trace and also to restore the continuity of the trace in the places where the ink is faint.

The droplet method, as described in this section, is an aggregate of ideas that existed in the research community under different forms [11, 12]. We provide a clear reference point here. In the present paper, our emphasis is on layout analysis of handwritten documents and we are interested in the ability of the droplet method to separate text lines with good preservation of the ascenders and descenders. In the digit recognition area, a related, but different, "droplet" technique is used to provide candidate segmentation points for touching digits.

Table 1: Performance of the KdK layout analyzer.

Layout element	Performance	Number of instances
page margins	1 error (0.1 %)	1040
rule lines of the table(color)	0 errors (0.0 %)	1040
rule lines of the table(grayscale)	9 errors (0.8 %)	1040
handwritten text lines	62 missed (0.2%) 173 overdetected (0.5%)	32816 31.6 per page
decision paragraphs	42 missed (0.6%) 280 overdetected (3,9%)	7130 6.9 per page

4. Results

The layout analysis system was applied on the document images of the KdK index from the year 1903. The results were saved in XML format.

Because we did not have ground-truth information, our layout analysis system was evaluated by manually checking the results and counting the errors. More detailed performance measures using the area overlap between predicted and target rectangles [5] were not applicable. Table 1 gives the numerical results. Our layout analyzer has a very good performance, the error rates being below 1% for finding most document elements. Only the (over)detection rate of the paragraphs requires significant further improvement. Nevertheless, on future scan sets, we can possibly use the layout analysis results automatically, without the tedious manual checking. The assumed risk is that a few errors will however be introduced.

Figure 4 shows some typical layout analysis results in which a colorization scheme was used to mark the different document regions. The same approach was also used for the manual checking in the evaluation of our system.

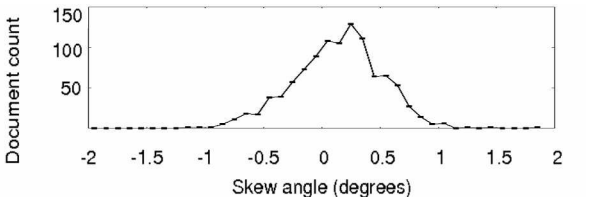

Figure 5: Skew distribution of the KdK documents contained in our experimental dataset (1040 scanned pages).

Figure 5 gives the histogram of the skew angles of the documents from our collection. Document skew was computed by linear regression on the bottom rule line of the index table. For almost all our documents, the

skew angle is less than 1 degree. And we decided to take no provisions regarding document skew in our present system. Figure 6 shows examples of line separators generated by the droplet method. The integrity of handwriting seems to be well preserved in most cases. Nevertheless, the merits of this line segmentation technique remain to be evaluated in the context of the complete retrieval system.

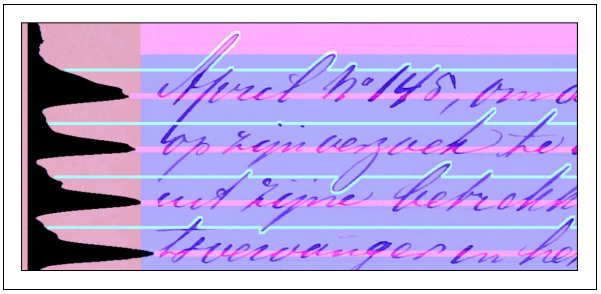

Figure 6: Examples of curvilinear separation paths generated by the droplet method.

The next stage in the development of our document retrieval system is to build specialized recognition modules on the basis of the document structure determined by the layout analysis. We have constructed a simple and effective recognizer for the machine-printed page numbers. They are located above one corner of the index table (see Fig. 4) and are extensively used for referencing throughout the KdK collection. The individual digits are extracted using vertical projection profiles and a rough estimate of the digit width (see Fig. 7). After normalization to 28x28 bitmaps, we applied nearest-neighbor classification using direct image matching. In leave-one-out tests, we obtained an excellent performance of 99.5% correct results (15 errors out of the 3053 digits extracted from the documents). These results are very promising because the same machine font was used for printing the page numbers over large portions of the KdK collection.

Figure 7: Two examples of machine-printed page numbers. The digits are often broken into several connected components (marked here using different colors / gray intensities). Vertical projections (shown below the numbers) together with an estimate of digit width are used to separate the individual digits and extract them for recognition.

5. Conclusions

Layout analysis is the first step towards the textual content of our documents and it becomes possible now to develop and apply dedicated recognizers on the separate logical document regions: month recognizer, day / number / digit recognizer, page title recognizer.

Our very good results are explainable by two factors:

- 1) the rigid structure of the KdK index allowed us to use top-down information in the design of the system and in the final checking of the layout results;
- 2) the right design choice of simple and robust image processing techniques provided the fundamental basis for building up the system. We feel confident that the largest part of the KdK collection spanning two centuries can be processed in a similar manner.

Layout analysis of handwritten documents remains a difficult open problem. In the current paper, we have presented a study-case that is relevant for the research projects aiming at digitizing large historical collections. The main challenge remains to build an effective word-spotting system for the main column of KdK collection where the textual content is essentially free. This is the focus of our current research efforts.

6. References

[1] NL-HaNA, 2.02.14, Archief van het Kabinet der Koningin, Den Haag (Netherlands), year 1903.

[2] H. Baird, "Digital libraries and document image analysis", *Proc. of the 7th ICDAR*, 2003, pp. 2—14.

[3] R. Cattoni, T. Coianiz, S. Messelodi, C. M. Modena, "Geometric layout analysis techniques for document image understanding: A review" *Technical report, IRST*, Trento (Italy), 1998.

[4] Y. Li, and Y. Zheng, D. Doermann, S. Jaeger, "A new algorithm for detecting text line in handwritten documents." *Proc. of the 10th IWFHR*. 2006

[5] S. Mao, A. Rosenfeld, T. Kanungo, "Document structure analysis: a survey." *Document Recognition and Retrieval X*, 2003

[6] U. Marti, H. Bunke, "Text line segmentation and word recognition in a system for general writer independent handwriting recognition." *Proc. of the Sixth ICDAR*, Seattle (USA), September 2001, pp. 159-163.

[7] G. Nagy, "Twenty Years of Document Image Analysis in PAMI." *IEEE Transactions on PAMI*. (Jan. 2000), pp. 38-62.

[8] L. O'Gorman, "The Document Spectrum for Page Layout Analysis." *IEEE Transactions on PAMI*. (Nov. 1993), pp. 1162-1173.

[9] N. Otsu, "A threshold selection method from gray-level histograms." *IEEE Transactions on Systems, Man and Cybernetics*, 1979, pp. 62–66.

[10] F. Shafait, D. Keysers, T. M. Breuel, "Performance Comparison of Six Algorithms for Page Segmentation." *In 7th IAPR Workshop on Document Analysis Systems*, Springer, Nelson (New Zealand), Feb 2006, pp 368-379.

[11] F. Venturelli, M. Zs. Kovacs-V, "An unconstrained handwritten line segmentation technique", *Fifth IWFHR*, University of Essex (England), 1996, pp. 385-388.

[12] A. Zahour, B. Taconet, M. Mercy, S. Ramdane, "Arabic handwritten text-line extraction", *Proc.of the 6th ICDAR*, Washington, DC (USA), 2001, pp. 281–285.