# Text Block Recognition in Multi-Oriented Handwritten Documents

Rainer Herzog, Arved Solth and Bernd Neumann
Department of Informatics
University of Hamburg
Germany
{herzog, solth, neumann}@informatik.uni-hamburg.de

***Abstract*—Automatic detection of text blocks is an important step before applying OCR or word-spotting techniques to document images. Our approach focusses on handwritten (historical) documents and uses the Gabor Transformation to facilitate this task. Apart from the main text, which often consists of rectangular shaped text blocks, marginalia are of special interest here. These areas are generally unconstrained regarding size, dimensions or orientation. Our system detects text blocks of at least three lines, representing a moderately homogeneous region regarding orientation and distances of text lines. Experiments on 40 documents, written in different european and asian writing systems, show good results, depending on the complexity of the layout.**

***Keywords*-document layout analysis; manuscript; text block recognition; Gabor Transform**

## I. Introduction

Manuscripts stored in European libraries have been studied for many decades, but vast amounts of manuscripts are still not yet completely categorized and remain nearly unknown. Sometimes bindings of books get destroyed, causing loose pages to trail away. One way to recover connections between physically separated manuscript parts is to analyse and compare their respective layouts, i.e. the outer dimensions of individual pages or text blocks, the number of text blocks contained on each page or the number of lines within each text block. Comparing these properties is one way to get first assumptions about the origin of manuscripts. Using image processing methods on digital images of manuscript pages, these measurements can be computed automatically and efficiently for large data sets.

Another target that will be addressed by our approach is skew correction of individual text blocks, for examples of glosses commenting the main text. The orientation of text lines within these glosses often varies considerably, particularly in Arabic manuscripts, Figure 1 shows a typical example. After estimating skew and line distance, skew correction can be done to allow further processing steps like word-spotting or OCR; this way these methods do not need to be scale or rotation invariant.

## II. Related Work

Most work done in the field of layout analysis aims at supporting the challenging tasks of word spotting and OCR. In 1985 Postl published an early paper about automatic layout analysis using Fourier analysis and simulated skew scans to detect the orientation of skewed lines in scanned documents [2]. As the described approach focuses on printed documents, it does not consider the specific problems that emerge when dealing with handwritten text.

Fig. 1. Example of an Arabic manuscript with diverse text blocks (Manuscript Ms.or.oct.261, p13v, Staatsbibliothek Berlin [1])

A similar approach was pursued by Jain and Bhattacharjee in 1992 using Gabor Filters [3]. They describe the problem of line detection in scanned images as a texture segmentation task. As in [2], the algorithm seems to have been applied only to printed documents, thus ignoring additional challenges of handwritten documents like irregular line orientations and distances.

Grouping handwritten text lines with similar orientation into zones and subsequent text line extraction is performed by Ouwayed and Belaid [4] using the Wigner-Ville Distribution (WVD) and connected component analysis. The WVD works similar to pixel projections for locating lines and spaces between lines. They evaluated their algorithm on 50 documents and achieved an accuracy of 97%. Their algorithms work

by first segmenting background from foreground, partitioning the input image, detecting multi-skewed zones and finally extracting individual text lines by following baselines.

In [5] Bukhari et al. propose to facilitate layout analysis of handwritten script using machine learning techniques. The goal of their work is to identify the main text block in a page containing complex layout and text blocks with different line orientations. In contrast to our approach they are not interested in distinguishing between individual text blocks apart from the main text area. The same target is addressed by Asi [6] with improved results. In a first step, a Gabor filter is used to coarsely segment the main text area. Then an energy function is defined based on the properties of the text components. A refined segmentation is then conducted by using graph cuts.

Garz et al. also tackle layout analysis in historic manuscripts [7]. The technique they describe is based on scale invariant feature transform (SIFT) and aims at identifying individual regions that are characteristic for the handwritings in question, like initials, headlines and text areas.

An approach utilizing contour detection to segment handwritten documents into individual lines is proposed by Bulacu et al. in [8]. They focus on the archive of the cabinet of the Dutch Queen (KdK) which contains a large set of handwritten historical documents. Since the project deals mainly with text written in Latin cursive script, it ignores characteristic challenges that arise when dealing with manuscripts using "unconventional" complex layout styles.

Our approach, presented in the subsequent sections, is new in several respects. First, it allows to delineate text blocks of arbitrary shape and orientation. Second, segmentation is based on both orientation and line distance discontinuities, resulting in block boundaries similar to those perceived by humans.

## III. Using the Gabor Transform for Layout Analysis

### A. Overview

The main characteristic property of a text block is the approximately constant periodicity of its line structure, both regarding orientation and line distance. Hence a Fourier Transform (FT) is in principle suited to detect such constancies, showing peak responses for pronounced spatial frequencies. Since we are interested in the local boundaries of text blocks, however, we apply a local frequency analysis in a sliding window approach. For each window position, the greyvalues are weighted by a Gaussian centered over the window and then subjected to a FT - this is known as a Gabor Transform (GT). The GT is an optimal short-time FT in the sense that it provides the maximal simultaneous resolution of a restricted image region and frequency band.

In order to cover all frequencies which may be relevant for the layout analysis, the GT is performed for several window sizes with Gaussians of varying width. In our approach, we accomplish this by applying a GT with fixed window size to a resolution pyramid, this way reducing the computational load. For each window position, the results of the GT at all resolution levels are analyzed for a peak magnitude corresponding to a periodic line structure. The criteria for accepting a local maximum are chosen to exclude peaks at high frequencies which may have been caused by the character structure within a text line.

The final step is image segmentation based on the local maximum obtained for each location. Discontinuities of both, frequency amplitude and orientation, are independent indicators of text block boundaries. After combining the gradients of these indicators in a single gradient image, segmentation is performed using the Watershed segmentation algorithm.

In the following subsections, we present details for these processing steps and motivate some of the parameter choices.

### B. Parametrizing the Gabor Transform and the Resolution Pyramid

The range of frequencies which can possibly correspond to a line structure in a manuscript is quite large. As an upper limit, we assume that the distance between two lines of text should be at least 10 pixels, corresponding to a character height of somewhat less. For a lower limit, we assume that the image contains at least 8 lines. The width of the GT in terms of the standard deviation $\sigma$ of the Gaussian has to be a compromise between the reliability of the frequency analysis and the spatial resolution. In our experiments, results were best for $\sigma$ = twice the smallest line distance, hence we have chosen $\sigma$ = 20. For ease of processing, windows are rectangular and limited to the range -3$\sigma$ ... +3$\sigma$ in both spatial dimensions, i.e. the window size is 121x121. The responses of a GT are quite selective in the frequency domain. Out of the 60 frequencies of a FT of this window size, we concentrate only on the most reliable wave numbers 8-12. Smaller wave numbers are increasingly influenced by noise or the bell shape of the Gaussian. Larger wave numbers correspond to periods supported by less and less pixels. In order to cover all frequencies which can be expected in a manuscript, the frequency analysis has to be repeated for window sizes increasing by a factor of 1.5, resulting in the useful frequency ranges 8-12, 12-18, 18-27 etc. As pointed out above, instead of increasing the window size, a more efficient method is to repeatedly reduce the image resolution by a factor of 1.5, creating a resolution pyramid. Hence for each window at each resolution level, the discrete GT is computed with constant parameters:

$$F_{k,l} = \sum_{m=0}^{120} \sum_{n=0}^{120} f[m,n] \cdot e^{-\frac{(n-60)^2+(m-60)^2}{800}} \cdot e^{-j2\pi(km+ln)/120}$$

### C. Performance for Artificial Data

In this subsection, we investigate and illustrate the performance of the GT for artificial grey value data. Figure 2 shows a GT of an artificial grey value distribution (red graph on top). *Zone A* represents a projection of some text lines, simulated as noisy sinusoid. In order to illustrate the effect of different window sizes, the Gaussian windows are set to $\sigma = 2.5\pi$ (green curve) and $\sigma = 1.25\pi$ (blue curve), where $2\pi$ is the wavelength of the sinusoid. These windows are shifted horizontally across the plot. The heatmaps on the

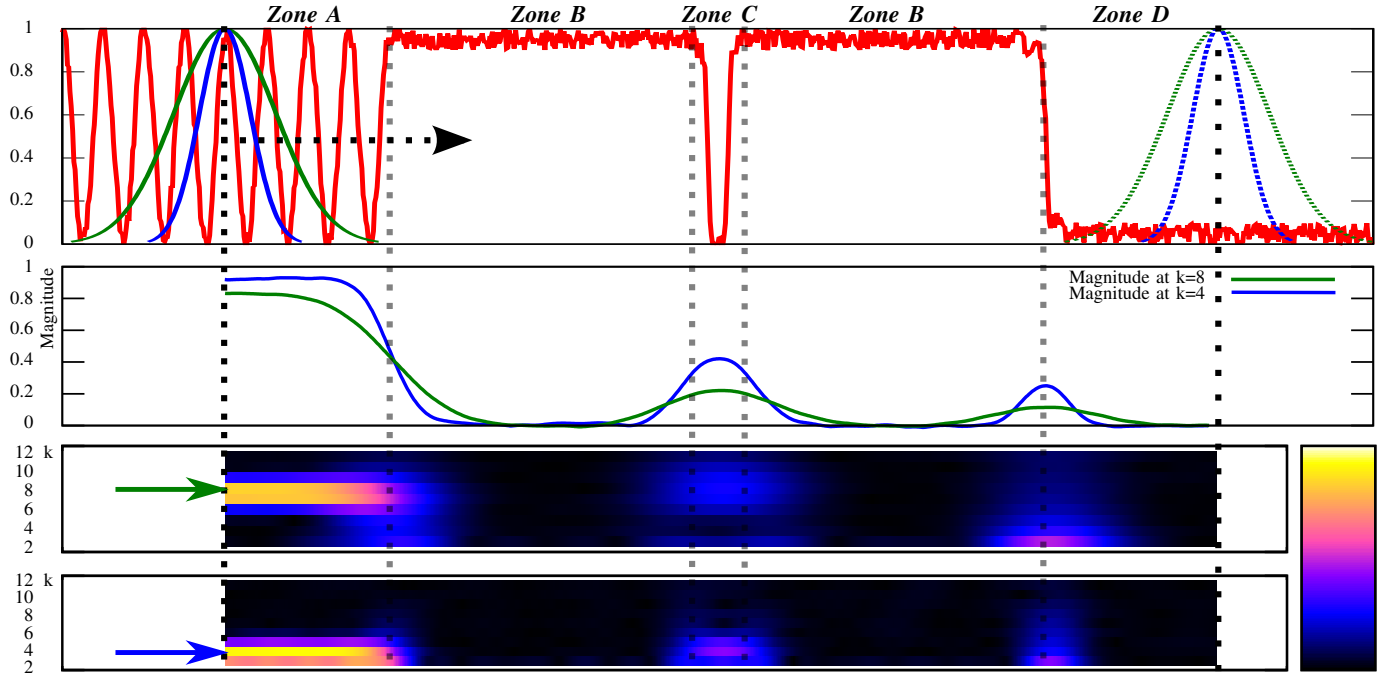
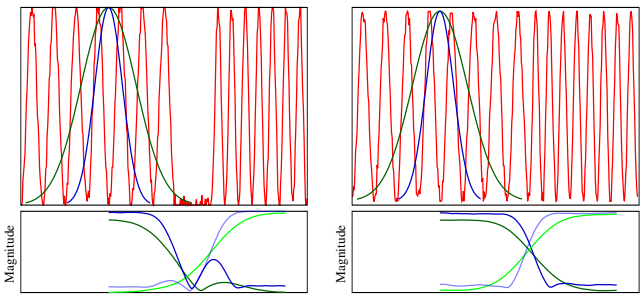
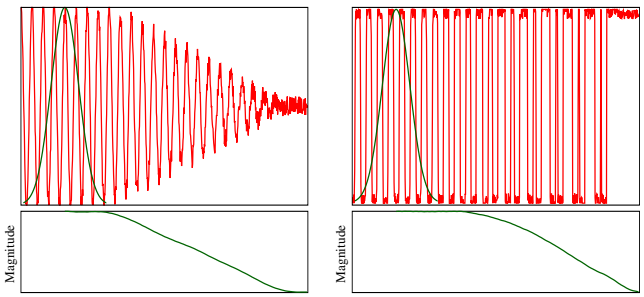



Fig. 2. 1D example of a GT applied to an artificial grey value distribution. See text for explanation




(a) Text blocks separated by gap (b) Touching text blocks

Fig. 3. 1D example of a GT applied to a projection of different line distances




(a) Decreasing contrast (b) Irregular line distance

Fig. 4. Deviations from idealised line projection representation lead to decreasing magnitude.

bottom show local maxima at k=8 for the green window and at k=4 for the blue. Plots of the distributions of these wave numbers are shown in between. *Zone B* represents a blank, nearly white area. The most interesting part is the transition from *Zone A* to *Zone B*. At the borderline, we see that the magnitude of the frequency is about half of the magnitude within the assumed text block. The smaller window reacts with a steeper declination, so we expect to get a more accurate result regarding the local extension of text blocks. However, the smaller window is more prone to overreact on noise like in *Zone C*, which could depict a thick line of a frame. Finally, *Zone D* represents a dark background often found on outer areas of document images. Here, the single transition from *Zone B* to *Zone D* causes a noticeable response. It is therefore useful to remove dark borders of images in advance by appropriate preprocessing.

Figure 3 illustrates transitions between blocks of different line distances. The darker blue and green plots represent a wave number optimal for the left side, the lighter ones are optimal for the right side. A smoother plot of the response of the best matching wave numbers is generated for the broader window (green plots). The blue plots for the smaller window react faster to changes of frequencies, however, they are more prone to be influenced by undesired interfering effects.

Figure 4 illustrates several other irregularities which influence the response of the operator. In real manuscript images, these may be due to variations of the ink colour or writing support, changing relations between script height and line spacings, and other deviations from an ideal block structure. Figure 5 shows an application of the operator to a section of a Tamil manuscript, demonstrating the combined effects of different line distances and a low contrast within the projection.

### *D. Determining Relevant Local Maxima*

The artificial examples support our decision to discern between two frequency bands. The “lower band”, comprising

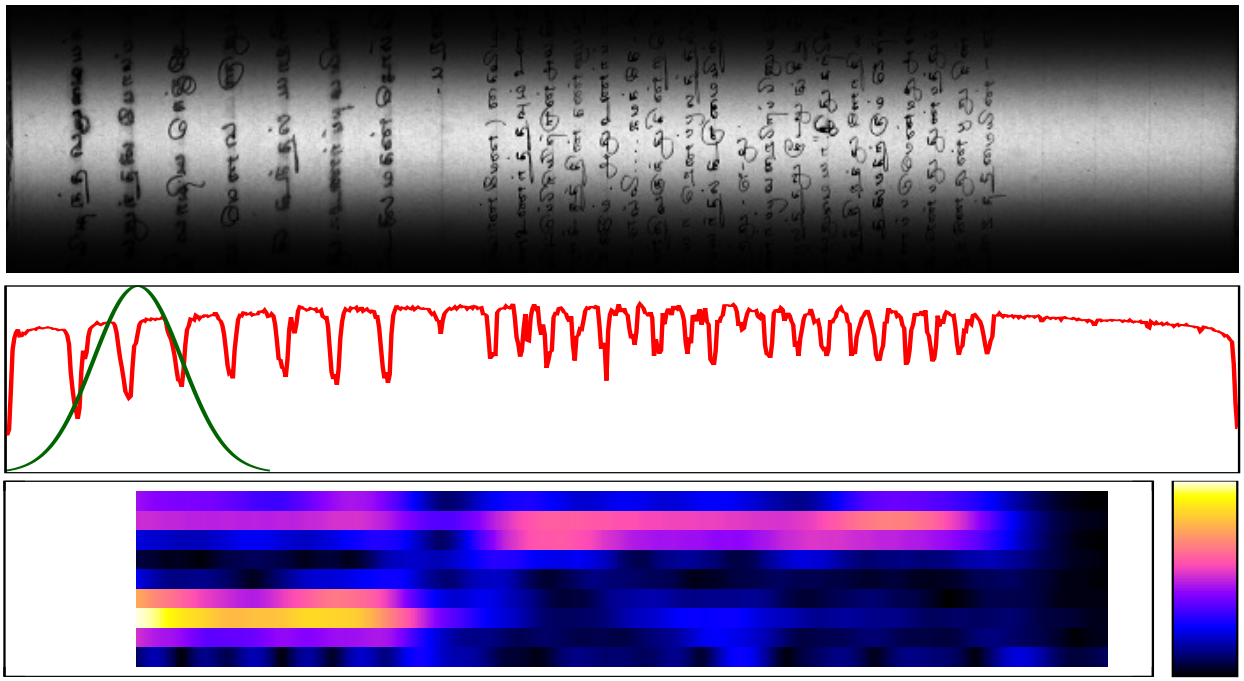


Fig. 5. Processing of a Tamil Manuscript

wave numbers 4-7, reveals a better spatial resolution, whereas the "upper band", consisting of wave numbers 8-12, has the advantage of more reliable magnitudes.

Unfortunately, the highest response obtained by the GT operator at a particular location and evaluated over all resolution levels does not always correspond to the relevant line frequency. Often, highest responses arise at frequencies at the character range level, as shown in Figure 6. In order to safeguard against this effect, we start the search for a maximum at the lowest resolution. For each maximum determined at one resolution level, we check the response for the same frequency at the next higher resolution level. Hence at one level, the maximum is located in the upper band, at the next level, this corresponds to a frequency in the lower band. We accept a maximum if the response in the lower band is more than 75% of the response in the upper band, thus making use of the advantages of both bands. If a maximum has been accepted, we disregard all further maxima at higher resolutions, if they are lower than twice of the actual one. This way, we prevent being mislead by meaningless maxima. Furthermore, a threshold is determined which discerns between regions containing text and spacings (shown in Figure 1 as shaded regions). This threshold can be adjusted by user interaction. We disregard the threshold in the following steps.

The whole process is illustrated for a challenging exemplar of an Arabic manuscript shown in Figure 1. Figure 7 shows the magnitudes of the resolution pyramid. All images are resampled to the size of the leftmost one with the highest resolution. The brightness encodes the maximum found for each pixel location.

Fig. 6. Strokes of characters with parallel orientation might lead to high responses of the corresponding frequency

### *E. Determining Text Block Boundaries*

The maximum determined for each pixel location indicates a specific frequency and orientation for the corresponding text block. In order to determine regions of homogeneous frequency and orientation, a gradient image is compiled by combining the gradient information of the frequency and the components of the normalized orientation vector. To allow upside-down orientation matches, the orientation angle is first doubled. The magnitudes of the frequency gradient image and of the gradient images of the components of the (recoded) orientation vector are then summed after a normalization.

Figure 8a shows the gradient image representing orientation discontinuities, Figure 8b shows the gradient of the frequency magnitudes, representing discontinuities of line distances, and Figure 8c shows the image of the combined gradients. This is then used as input for a watershed segmentation, typically leading to many small regions. In a subsequent step, these regions are merged until a threshold is reached, which can be set by the user (1/100 of the size of the text area determined in the first step is a good starting point). The resulting 8 regions are shown in Figure 8d, numerical values for region properties are presented in Table I. Due to the inhomogeneity within text Region 1 featuring interlinear glosses, Region 8

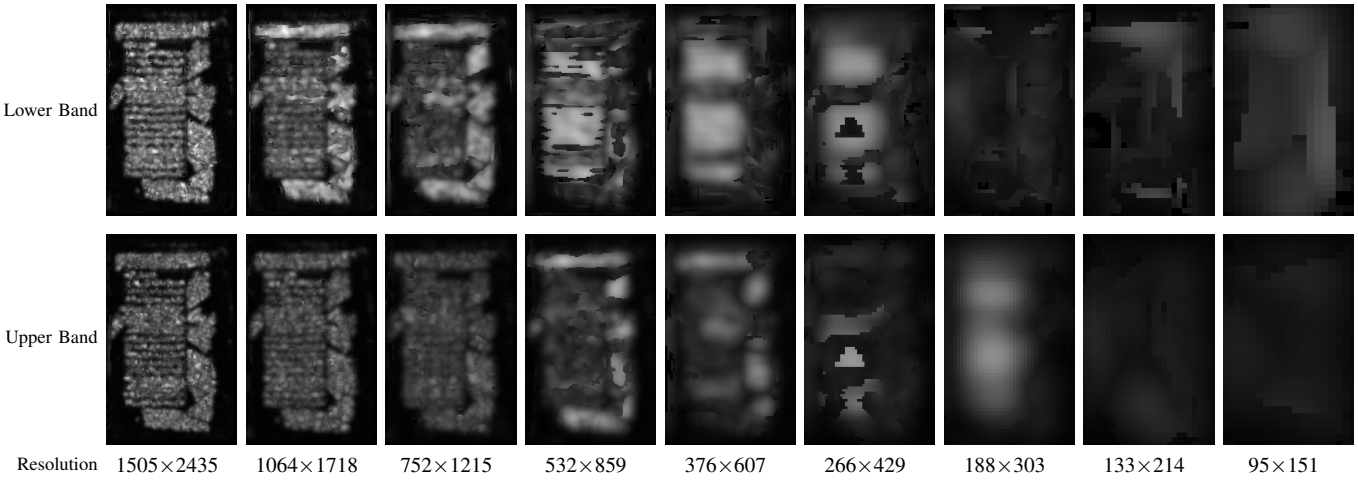
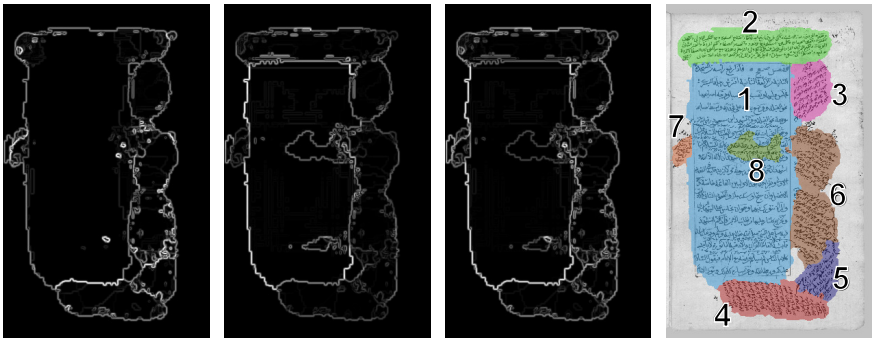



Fig. 7. Resolution pyramid: Magnitudes of frequencies at specified resolution.




(a) Orientation gradient (b) Frequency gradient (c) Combined gradients (d) Final regions

Fig. 8. Determining homogeneous text block areas.

was identified as individual region. If a priori knowledge is available, inclusions within the main text blocks showing a higher frequency can be filtered out easily. In contrast, the text blocks of Region 6, written in similar orientation and line spacing, are merged because of too little space between them.

TABLE I
PROPERTIES OF REGIONS FROM FIGURE 8D

| | COG (x,y) | Area [Pixel] | Line distance mean [Pixel] | Line distance variance | Line orientation mean [0-180] |
|---|---|---|---|---|---|
| **Region 1** | (545,1220) | 45623 | 89.93 | 77.777 | 90.39 |
| **Region 2** | (645,300) | 10421 | 44.95 | 221.5 | 93.25 |
| **Region 3** | (1050,620) | 4412 | 45.78 | 85.6 | 29.58 |
| **Region 4** | (775,2150) | 7108 | 35.35 | 60.97 | 164.12 |
| **Region 5** | (1140,1965) | 3247 | 39.55 | 69.22 | 33.62 |
| **Region 6** | (1075,1380) | 10920 | 40.67 | 112.92 | 114.88 |
| **Region 7** | (115,1070) | 953 | 36.65 | 192.38 | 98.95 |
| **Region 8** | (655,1045) | 2226 | 49.53 | 133.38 | 90.43 |

## IV. EXPERIMENTAL RESULTS

In further experiments, we aim at detecting text blocks of at least three consecutive lines of text, written in about the same distance and orientation. Text areas consisting of only two lines are ignored, hence it is not considered as an error if these areas are not detected.

### A. Test Set

To test our method, 40 different manuscript images were selected, the links to all images can be found here [9]. They mainly contain non-curved text blocks, so that it is possible to measure an average orientation of the blocks. The 40 images are subclassified into 5 different test sets:

- In the ***Berlin*** set, 15 images stem from seven different Arabic manuscripts [1]. All of them consist of one main text block and between 3 and 18 paratext blocks (on average 7), and highly different orientations.
- The ***Leipzig*** set has similar properties and comprises another 10 images from two different Arabic manuscripts [10]. They embody between 2 and 18 paratext blocks, 9 on average, next to the main text block.
- We also tested 10 manuscript images from four different (***Greek***) manuscripts [11], which represent, with one exception, only horizontally written text. Five manuscripts consist of one main text block and up to four paratext blocks, another five contain up to four text blocks without a prominent main text block.
- Four (***Borno***) Qur'an manuscripts from Nigeria were used, consisting of one main text each in Arabic and

between 10 and 17 paratext blocks, written in Arabic and old Kanembu. The paratext blocks are written mostly in horizontal or vertical orientation.

- One poster (***Fol547b***) consisting of five different horizontally written text blocks in four different languages is taken from [1].

### B. Ground Truth

In order to provide a pixel-accurate ground truth, each image was binarised to separate the text from the background. Then, text areas consisting of at least three consecutive lines were identified. After manually measuring the average orientation and line distance of each block, a dilation operation was applied to all pixels of a single text block. A radius half of the average line distance was choosen to create closed regions representing the extent of each block.

### C. Detection Rate without Segmentation

To prove the general applicability of our approach, a pixelwise comparison of the measured orientation and line distance of all pixels representing ground-truth text blocks with the values determined for the window centered at the same position was conducted. These values can also be used to provide interactive information. If, for example, a user inspects a manuscript image and moves the cursor over some text blocks, orientation and line-distance at the position of the cursor can be displayed. Table II shows the amount of pixels matching the ground truth values within the given limits for each set of documents. If the documents comprise a main text block, a second result is given for all comparisons excluding pixels belonging to the main text area, which is a much more challenging task. The main text areas consisting of about 20 lines are normally written with considerable more uniformity regarding text height, orientation and line distance; in contrast, small paratext areas of only three lines cannot provide an equivalently strong signal due to their size, even if they were written very accurately. Very often, however, it is not even possible to provide a reliable ground truth with a stricter tolerance level than $\pm$ 5-10° regarding orientation or $\pm$ 5-10% regarding line distance, due to their sloppy writing style.

TABLE II
PIXEL-WISE COMPARISON OF GROUND TRUTH TEXT REGIONS WITH DETERMINED VALUES AT THE SAME POSITION

| Amount of pixels matching the ground truth | Difference of orientation up to | | | | | Difference of line distance up to | | | |
|---|---|---|---|---|---|---|---|---|---|
| | 5° | 10° | 15° | 20° | 30° | 5% | 10% | 20% | 30% |
| **Berlin**, all text areas | 83,1% | 92,7% | 95,4% | 96,4% | 97,5% | 72,2% | 85,2% | 90,5% | 92,1% |
| without main text area | 59,0% | 82,0% | 89,4% | 92,0% | 94,7% | 41,4% | 65,6% | 79,4% | 83,6% |
| **Leipzig**, all text areas | 79,7% | 92,0% | 96,2% | 97,2% | 97,7% | 75,7% | 90,6% | 95,2% | 95,8% |
| without main text area | 50,1% | 80,5% | 91,8% | 93,9% | 95,0% | 52,5% | 80,7% | 89,8% | 91,2% |
| **Greek**, all text areas | 96,6% | 98,8% | 99,3% | 99,5% | 99,5% | 70,5% | 88,6% | 94,7% | 95,6% |
| without main area | 92,0% | 96,6% | 98,1% | 98,8% | 98,8% | 50,8% | 70,0% | 85,2% | 88,9% |
| **Borno**, all text areas | 86,2% | 92,1% | 93,2% | 93,4% | 93,5% | 47,5% | 64,4% | 71,9% | 72,9% |
| without main text area | 72,2% | 83,9% | 85,9% | 86,3% | 86,6% | 29,0% | 48,3% | 61,4% | 63,8% |
| **Folio547b**, all text areas | 97,0% | 98,8% | 98,8% | 98,8% | 98,8% | 80,1% | 88,2% | 93,7% | 94,4% |

As Table II reveals, the recognition rates regarding the determination of the line distance of the *Leipzig* set are slightly better than for the *Berlin* set, as the paratext blocks of the *Leipzig* set are more uniformly written.. Regarding orientation,

TABLE III
MATCHING OF GROUND TRUTH TEXT REGIONS WITH DETERMINED REGIONS

| Matching text regions | | >90% | >80% | >70% | >50% |
|---|---|---|---|---|---|
| **Berlin** | Fraction of matched text regions | 7/122 | 30/122 | 48/122 | 69/122 |
| | Area of all matched text regions | 25,21% | 64,86% | 76,19% | 80,78% |
| | Difference of line distance | 1,70% | 3,48% | 7,64% | 13,75% |
| | Difference of orientation [°] | 1,0 | 2,8 | 2,7 | 9,5 |
| **Leipzig** | Fraction of matched text regions | 11/98 | 33/98 | 46/98 | 57/98 |
| | Area of all matched text regions | 35,26% | 67,19% | 76,61% | 80,93% |
| | Difference of line distance | 2,78% | 4,92% | 3,76% | 6,31% |
| | Difference of orientation [°] | 2,1 | 3,6 | 4,5 | 6,8 |
| **Greek** | Fraction of matched text regions | 4/31 | 9/31 | 14/31 | 18/31 |
| | Area of all matched text regions | 27,15% | 58,33% | 79,06% | 84,23% |
| | Difference of line distance | 0,88% | 4,89% | 4,49% | 3,66% |
| | Difference of orientation [°] | 0,1 | 0,3 | 0,6 | 0,7 |
| **Borno** | Fraction of matched text regions | 0/58 | 2/58 | 9/58 | 22/58 |
| | Area of all matched text regions | 0,00% | 3,40% | 17,18% | 39,49% |
| | Difference of line distance | - | 13,38% | 15,06% | 5,49% |
| | Difference of orientation [°] | - | 1,3 | 2,2 | 2,6 |
| **Folio547b** | Fraction of matched text regions | 1/5 | 2/5 | 2/5 | 2/5 |
| | Area of all matched text regions | 26,74% | 43,48% | 43,48% | 43,48% |
| | Difference of line distance | 0,38% | 1,07% | - | - |
| | Difference of orientation [°] | 0,2 | 0,0 | - | - |

| | Regions missed | Regions misdetected |
|---|---|---|
| **Berlin** | 1/122 | 9 |
| **Leipzig** | 2/98 | 1 |
| **Greek** | 1/31 | 8 |
| **Borno** | 0/58 | 0 |
| **Folio547b** | 0/5 | 1 |

the *Leipzig* and *Berlin* sets provide similar results. The *Greek* set predominantly consists of horizontally written text, the recognition results for orientation are nearly perfect. Regarding line distance, results are similar for the *Berlin* and *Leipzig* sets. Not surprisingly, the results of the *Borno* set are worst, as these documents are heavily degraded and excel in highly varying contrast, extremely densely written text with very little interlinear space, and much interlinear text within the main text areas. *Folio547b* yields nearly perfect results, as the four major text blocks are written very cleanly.

### D. Recognition of text regions

Table III presents the results of the recognition of text regions. According to the process described in Section III-E, regions are formed based on similar values regarding orientation and line distance. In contrast to other approaches, no assumptions about the characteristics of the documents or shapes of text blocks were made. The first row of each set in Table III states the fraction of recognized text blocks matching the corresponding ground-truth block. The criteria for a match is an overlap with at least the percentage shown at the top of the column, which has to be true for matching in both directions. The second line shows the percentage of text area pixels of the ground truth overlapping with a matching region. Here, the influence of the mismatch of small text areas is less apparent, and the match of the main text areas leads to a significant improvement.

In the first two rows of each evaluation, matches of text regions are only determined on the basis of overlapping areas. Consecutive columns include the results of the previous column. The following two rows compare the average values of orientation and line distance between detected and ground truth regions. The data of these two rows is separately computed for each column.

The second part of the table addresses text areas of the

Fig. 9. Top and middle row: Samples of the *Berlin*, *Leipzig*, *Borno* and *Greek* sets. Automatically detected regions are shown in the top row, the ground truth is represented below. Bottom row: Automatic detection and ground truth of *Folio547b*. Different colours represent different text blocks.

ground truth, which are not covered by any determined area by at least 25% overlap and are therefore considered missed regions. Vice versa, detected regions are listed here that are not supported by at least 25% of any region of the ground truth and are therefore considered false detections. The reasons for misdetections of altogether 4 out of 319 blocks can be explained by insufficient contrast, as in the *Greek* set, or too small region size, both causing a signal below the threshold. Reasons for false detections (19 incidents altogether) are patterns like a watermark with regular structure within the *Greek* set or coincidentally occurring regular structures involving single-line text or stains.

Regarding the values for line distance and orientation, neither orientation nor line distance could be correctly determined for three textblocks of the *Berlin* set, consisting of three or four lines, due to an overly irregular line structure and some noise. Apart of these, the agreement is good for all regions which overlap by more than 70%, and even for areas with smaller area match. The main reason for the good conformance is the fact that quite often text blocks with similar line structures and orientation are placed in close distance and therefore not split up. This is especially striking in the case of *Folio547b*, where only the two text blocks at the left and right could be matched, while the two in the middle were not separated due to the conforming line structure, and therefore could not be matched, leading to a suboptimal recognition rate for this poster. The same is true for the *Greek* set, where most segmentation errors are caused by neighbouring text blocks of similar line distance and orientation.

The lack of white space areas in the *Borno* set makes processing of these documents extremely difficult. When establishing the ground truth, the separation of text blocks could in some cases only be identified by a change in the ink type, in the writing style or by a thin drawn line. The structure of most of the main text area could be determined for this set, though many paratext lines have been inserted between the main text lines. As the main text is written in a very dark ink and with thicker strokes, the structure of the main text block predominates the smaller structures, proving the robustness of the approach against significant noise.

The threshold which separates text from non-text regions, mentioned in Section III-D, was adjusted for only four of all forty documents.

Figure 9 presents one typical result for each set.

## V. Conclusion

In our paper we propose a binarization-free method for determining the characteristics of text blocks of any shape written in arbitrary orientations in handwritten documents. In a further step we show the segmentation of these text blocks, based only on computed orientation and line distance, without presuming special properties like the script size or shape of text blocks. Our approach is based on the Gabor Transform to identify text regions with roughly homogeneous line orientation and line distance. We have shown that our approach is robust enough to allow slight variations of orientation and line distance within individual text blocks. Our system is not limited to any particular kind of script and can be applied to arbitrary digital images containing text. Our segmentation process fails mainly when text blocks with similar orientation and line distance touch. This could be improved by rotating questionable regions into a horizontal or vertical orientation, and then applying methods of the spatial domain, such as XY-cut. This would greatly improve the result on *Folio547b*.

## Acknowledgment

This work has been funded by the German Research Foundation (DFG) as collaborative research project SFB950 “Manuskriptkulturen in Asien, Afrika und Europa” aka “Centre for the Study of Manuscript Cultures” (CSMC).

## References

[1] Staatsbibliothek Berlin, “Datenbank der orientalischen Handschriften der Staatsbibliothek zu Berlin.” [Online]. Available: http://orient-digital.staatsbibliothek-berlin.de

[2] W. Postl, “Detection of Linear Oblique Structures and Skew Scan in Digitized Documents,” in *Pattern Recognition, International Conference on*, 1986, pp. 687–689.

[3] A. Jain and S. Bhattacharjee, “Text segmentation using gabor filters for automatic document processing,” *Machine Vision and Applications*, vol. 5, no. 3, pp. 169–184, 1992.

[4] N. Ouwayed and A. Belaïd, “Multi-oriented text line extraction from handwritten arabic documents,” in *Document Analysis Systems, 2008. DAS ’08. The Eighth IAPR International Workshop on*, Sept 2008, pp. 339–346.

[5] S. Bukhari, T. Breuel, A. Asi, and J. El-Sana, “Layout analysis for arabic historical document images using machine learning,” in *Frontiers in Handwriting Recognition (ICFHR), 2012 International Conference on*, Sept 2012, pp. 639–644.

[6] A. Asi, R. Cohen, K. Kedem, I. Dinstein, and J. El-Sana, “A coarse-to-fine approach for layout analysis of ancient manuscripts,” in *Frontiers in Handwriting Recognition (ICFHR), 2014 International Conference on*, Sept 2014.

[7] A. Garz, R. Sablatnig, and M. Diem, “Layout analysis for historical manuscripts using sift features,” in *Document Analysis and Recognition (ICDAR), 2011 International Conference on*, Sept 2011, pp. 508–512.

[8] M. Bulacu, R. van Koert, L. Schomaker, and T. van der Zant, “Layout analysis of handwritten historical documents for searching the archive of the cabinet of the dutch queen,” in *Document Analysis and Recognition, 2007. ICDAR 2007. Ninth International Conference on*, vol. 1, Sept 2007, pp. 357–361.

[9] Test set of 40 manuscript images. [Online]. Available: http://kogs-www.informatik.uni-hamburg.de/%7Eherzog/testset.txt

[10] Orientalisches Institut, Universität Leipzig, “Islamic Manuscripts at the Leipzig University Library.” [Online]. Available: http://www.islamic-manuscripts.net

[11] Greek State Archives, “The digital collections of the Greek State Archives.” [Online]. Available: http://arxeiomnimon.gak.gr/en/index.html

[12] S. Qian and D. Chen, “Discrete gabor transform,” *Signal Processing, IEEE Transactions on*, vol. 41, no. 7, pp. 2429–2438, Jul 1993.

[13] F. Harris, “On the use of windows for harmonic analysis with the discrete fourier transform,” *Proceedings of the IEEE*, vol. 66, no. 1, pp. 51–83, Jan 1978.

[14] B. Jähne, *Digitale Bildverarbeitung*. Springer, 2005.