Proceedings of the 38th Annual Conference on Computer Applications and Quantitative Methods in Archaeology, CAA2010
F. Contreras, M. Farjas and F.J. Melero (eds.)

Text Detection in Ancient Manuscripts Using Orientation- and Frequency-Signatures of the Texture

Garz, A. , Gau, M. , Sablatnig, R.
Computer Vision Lab, Institute of Computer Aided Automation, Vienna University of Technology, Austria
{garz, mgau, sab}@caa.tuwien.ac.at

Text detection in ancient documents faces specific challenges: Due to the age of the materials and inappropriate storing conditions, the documents are degraded and stained, as well as the ink is partially faded out . Furthermore, fluctuating text lines, skewed text blocks, superimpositon of text elements – such as initials – or varying layouts are problems to cope with. Hence, a texture-based approach which exploits the fact that different kinds of textures have dissimilar distributions of orientations and frequencies is proposed. The main idea is to exploit the properties of the Auto-Correlation Function (ACF) to classify document areas. The ACF is computed in a local region with the aid of sliding windows of different sizes. Support Vector Machines are used to perform the classification task. The approach proposed is applied to Glagolitic manuscripts from the 11th century. The evaluation is based on manually labeled ground truth and shows the accuracy of the features chosen, even when the method is applied to document pages that are different in writing style and line spacing to those in the training set.

Keywords: layout analysis, texture, ancient manuscripts.

1. Introduction

This paper presents a texture-based approach for text area detection in ancient documents. The approach introduced is used for text detection in the Old Church Slavonic Psalter of Demetrius (*Cod. Sin slav. 3/N*), which is a manuscript written on parchment in Glagolitic letters. It was found in 1975 in St. Catherine's monastery on Mt. Sinai, Egypt, and probably originates from the 11th century (MIKLAS *et al*., 2008). The Psalter comprises 145 folios, each of which consists of two pages: recto (r) and verso (v).

Page and text layout analysis constitutes an fundamental part in the study of ancient manuscripts. Especially in the case of degraded material like our sample of a partly faded, chipped, blurred and stained medieval parchment codex, layout segmentation aids philological examination and text decipherment. Detecting illustrations, headlines and textual decorations and recognizing the layout patterns, reveals information about the genesis of the object, in particular on its scriptorium, the textual tradition, as well as its spatio-temporal origin. Furthermore the computational purpose of document layout analysis is to provide a preparatory stage for automated investigations. Splitting a document into homogeneous elements, such as text regions or images allows layout area classification (OKUN and PIETIKÄINEN, 2000). Finding text areas can be used as a pre-processing step for other algorithms like text reconstruction, text line segmentation (KLEBER *et al*., 2008b) and Optical Character Recognition (DIEM and SABLATNIG, 2009).

Traditional binarization-based approaches developed for layout analysis of machine-printed documents cannot be applied efficiently to ancient manuscripts due to material, the age of documents, non-appropriate storing conditions or degradation processes (KLEBER *et al*., 2008a). Therefore, a texture-based method is used to manage these problems.

Furthermore, handwritten documents have varying layouts, fluctuating text lines, non-constant spacing between words and lines or superimposing of text ele-ments (LIKFORMAN-SULEM *et al*., 2007). Hence, texture-based methods that imply rectangular layouts or machine-printed text are not applicable.

Figure 1 shows three folios of the Psalter that present the following problems: in all folios, varying line spacing and fluctuating text lines appear. The parchment of folio 51v is decayed and stained, and characters are written between the lines. Folio 106r comprises different writing styles and line spacing; the parchment is corrugated. In folio 140r, the ink is faded out, the pores of the parchment emerge, and the text is partially overwritten with a different writing utensil.

Figure 1: *Three pages of the Old Church Slavonic Psalter of Demetrius.*

In (BULACU *et al.*, 2007), Projection Profiles (PP) based on the number of transitions between ink and paper are used as the main analysis method for structured manuscripts. The authors of (ACHARYYA and KUNDU, 2002) implement a multi-scale wavelet analysis to extract text regions of machine-printed documents. A Gabor Function based filter bank is proposed in (RAJU *et al.*, 2004) to extract text of machine-printed documents based on spatial frequency properties of the text. In (JOURNET *et al.*, 2008) a multi-scale approach based on the Auto-Correlation Function (ACF) is proposed to characterize the content of documents from the 15th and 16th century.

Due to problems of unstructured manuscripts men-tioned before, PP as only method for text detection would fail for these documents (BULACU *et al.*, 2007). Wavelets and Gabor filters are methods to extract frequencies and orientations of textures. However, authors in (JOURNET *et al.*, 2008) have shown that when applied on documents that contain drop caps and drawings their ACF-based approach performs better than the Gabor filters.

The method presented in this paper detects written areas by studying the orientations and frequencies that are present in the image. Text regions are strongly correlated in writing direction as they are organized in lines, even when text blocks are skewed or characters have different sizes. Therefore, the ACF is chosen as the principal concept. The approach is applied to three image scales with the help of sliding windows.

This paper is organized as follows. In the subsequent section the proposed method is described. Then, results are presented, followed by a conclusion.

2. Methodology

The algorithm developed draws its main contribution from (JOURNET *et al.*, 2008). It has three main steps: first, pre-processing is done, then features based on the ACF and ink-paper transitions are extracted, and finally, classification is performed with Support Vector Machines (SVM).

Due to the fact that the color of the background (gray) and the parchment (yellowish to brown) in the dataset used are distinguishable, a color-histogram-based segmentation method is used to separate the page from the background.

2.1. Feature Extraction

Features linked to orientations and frequencies are computed locally at three different scales to analyze the texture. Frequency in this context means transitions between ink and parchment. A total of three different feature vectors are used to classify an image region. Two feature vectors are computed from the ACF's response. The third feature vector used is based on frequencies presented in the image.

As using a multi-scale technique permits identifying structures of various scales (JOURNET *et al.*, 2008), features are computed at three image scales with the aid of sliding windows with different sizes.

The ACF makes it possible to examine the directions existing in an image. On the ACF, *"the data related to a same direction will be located on a same straight line"* (JOURNET *et al.*, 2008). A tool called Rose of Directions (ROD) is used to study the response of the ACF, as proposed in (JOURNET *et al.*, 2008). This diagram is computed as the polar transformation of the AFC's output. In Figure 2, two different textures, namely background in the first and text in the second row, the

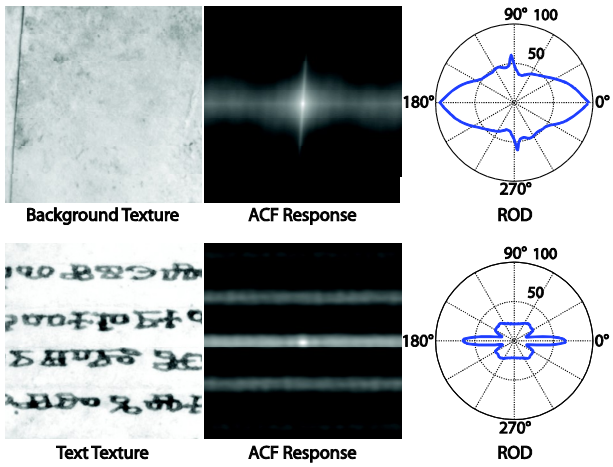

Figure 2: *Texture, corresponding ACF and ROD.*

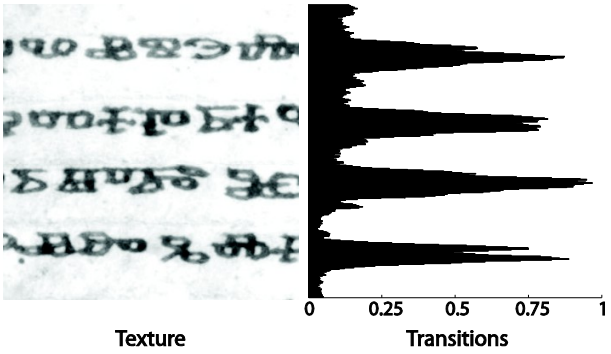

Figure 3: *Window of text and the corresponding transition histogram.*

corresponding response of the ACF and the ROD are shown.

For text regions, the shape of the diagram is characteristic. It has two peaks in the writing direction and a blob in the middle. The exact shape depends on three factors: the character-size, the orientation and the number of the text lines (JOURNET *et al*., 2008). This is not true for non-text areas where the ROD has an arbitrary shape.

The first feature vector extracted consists of three values proposed by JOURNET *et al*. in (JOURNET *et al*., 2008). First, the angle of the main direction is used which is normalized to 0–180° as the deviation of the horizontal angle. The second value is the intensity of the ACF which corresponds to the maximum value of the ROD. The last feature gives a measure for the overall shape of the ROD and is calculated as the Standard Deviation σ. A ROD with several strong orientations leads to a high σ.

For the second feature vector – ROD histogram – the ROD is shifted such that the main orientation is at 0° to be able to compare different RODs and to achieve invariance to skewed text lines. As the ROD is symmetric, only 0–180° are taken for the calculation of the ROD histogram feature vector. To further decrease the order of the dimensionality, the number of features is down-sampled to half. Experiments have shown that this reduction of dimensionality does not impair the classification performance.

Finally, the third feature vector is computed as a transition histogram. Transitions are computed as the first derivation of the window in horizontal direction (BULACU *et al*., 2007) which is then accumulated horizontally and normalized. This leads to a projection profile which has peaks at locations of text lines (BULACU *et al*., 2007). Figure 3 shows a window containing text and the corresponding transition histogram.

3. Classification

Having computed the features, classification is performed by multiple SVMs. For every feature vector and window size respectively, a SVM using a Radial Basis Function as kernel is trained independently. Then, the individual classification results are combined through class probabilities.

4. Results

The proposed method is applied to a random sample of 90 folios of the Psalter which show varying layouts and writing styles. The training set contains four different pages. The evaluation is based on manually labeled ground truth. One exemplified page with labeled ground truth is shown in Figure 4 on the left, while the classification result of the algorithm proposed in this paper is given in the right column. Areas surrounded by a black line are text regions that were detected. As can be seen, initials (on the left) and headlines (in the center of the page, indicated by larger characters than in the regular text) are not labeled in the ground truth.

The performance of the methodology is evaluated using the F-score (VAN RIJSBERGEN, 1979). For this test setting, an average F-score of 0.937 (with a standard deviation of 0.03) is reached. The precision is at 0.908 whereas the recall at 0.906.

On average, 3.17 % of the pixels that belong to text regions are not identified as text (False Negatives, FN) whereas background-pixels wrongly classified as text (False Positives, FP) occur at 3.31 %. For this application it has to be pointed out that FP have less impact than text that is not detected because FP are, for example, gaps between text lines that occur when a paragraph does not justify to the right text border, large line spacing or headlines classified as text (see Figure 4).

Figure 5 presents three classification results that emphasize characteristics of the approach proposed. Areas surrounded by black lines are text regions. Single lines, first and last lines of a page are not reliably detected by the ACF-based method. This is because of the properties of the ACF mentioned earlier. Headlines that are charactterized by taller characters and a different aspect ratio may be classified as text as well (see Figure 4 – folio 74r). Text areas that have very small or big line spacing – such as in folio 106r (see Figure 5) – have a

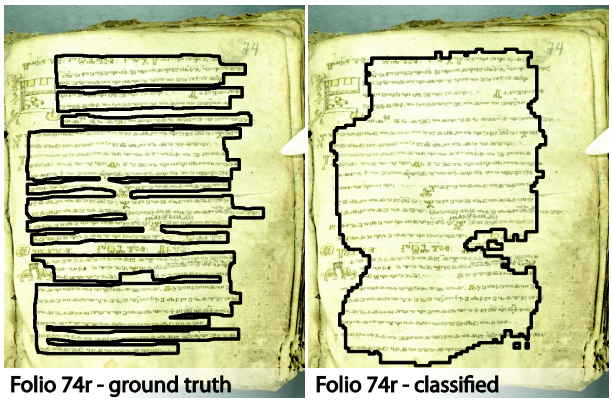

Figure 4: *Folio 74r with manually labeled ground truth and classified by the method.*

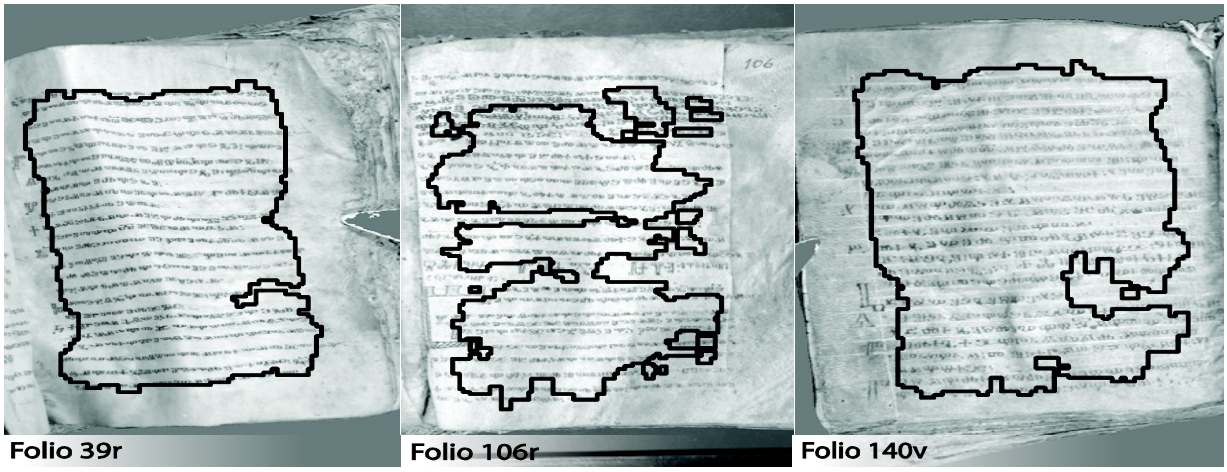

Figure 5: *Classified folios.*

different texture and are therefore not reliably detected. The method is tolerant towards varying character sizes (see Figure 5), skewed text blocks (see Figure 5 – folio 39r) and faded-out ink (see Figure 5 – folio 140v).

Conclusion

An approach on the basis of texture-analysis is pro-posed which is performed on three scales with the aid of sliding windows of different sizes. The main concept is to use characteristics of the ACF to detect text areas in ancient degraded manuscript images. In addition, a vertical profile of transitions between ink and parchment is used. Classification is performed with SVMs.

The evaluation shows the accuracy of the method chosen for text detection in ancient documents with heterogeneous background, even when the method is applied to document pages that are different in writing style and line spacing in comparison to those in the training set. The classification performs well even with a training set consisting of four samples.

Future work will include the use of a larger training set, of at least 20 to 30 training images. Furthermore, an improved rating algorithm for the combination of the individual classification results of the SVMs is needed. Therefore, this algorithm will take into account the reliability of each feature vector. In addition, a rating based on the window sizes is in the scope of future work.

Acknowledgements

This work was supported by the Austrian Science Fund under grant P19608-G12.

References

ACHARYYA, M., and KUNDU, M. K., 2002. Document Image Segmentation Using Wavelet Scale-Space Features. *IEEE Trans. on Circuits and Systems for Video Technology , Vol. 12,* pp. 1117-1127.

BULACU, M., VAN KOERT, R., SCHOMAKER, L., and VAN DER ZANT, T., 2007. Layout Analysis of Handwritten Historical Documents for Searching the Archive of the Cabinet of the Dutch Queen. *ICDAR, Vol. 1*, pp. 357-361

DIEM, M., and SABLATNIG, R., 2009. Recognition of Degraded Handwritten Characters Using Local Features. *ICDAR,* pp. 221-225.

JOURNET, N., MULLOT, R., EGLIN, V., and RAMEL, J.-Y., 2008. Analyse d'Images de Documents Anciens: une Approche Texture. *Revue Traitement du Signal (Presse univ. de Grenoble) , Vol. 24*, pp. 461-479.

JOURNET, N., RAMEL, J.-Y., MULLOT, R., and EGLIN, V., 2008. Document Image Characterization Using a Multiresolution Analysis of the Texture: Application to Old Documents. *IJDAR , Vol. 11*, pp. 9-18.

KENNARD, D. J., and BARRETT, W. A., 2006. Separating Lines of Text in Free-Form Handwritten Historical Documents. *DIAL*, pp. 12-23.

KLEBER, F., SABLATNIG, R., GAU, M., and MIKLAS, H., 2008a. Ancient Document Analysis Based on Text Line Extraction. *ICPR*, pp. 1-4.

KLEBER, F., SABLATNIG, R., GAU, M., and MIKLAS, H., 2008b. Ruling Estimation for Degraded Ancient Documents Based on Text Line Extraction. *EVA*, pp. 79-86.

LIKFORMAN-SULEM, L., ZAHOUR, A., and TACONET, B., 2007. Text Line Segmentation of Historical Documents: a Survey. *IJDAR , 9*, pp. 123-138.

MIKLAS, H., GAU, M., KLEBER, F., DIEM, M., LETTNER, M., VILL, M., ET AL., 2008. Slovo: Towards a Digital Library of South Slavic Manuscripts. Sofia: Boyan Penev.

OKUN, O., and PIETIKÄINEN, M., 2000. A Survey of Texture-Based Methods for Document Layout Analysis. *WTAMV*, pp. 137-148.

RAJU, S. S., PATI, P. B., and RAMAKRISHNAN, A. G., 2004. Gabor Filter Based Block Energy Analysis for Text Extraction from Digital Document Images. *DIAL*, p. 233+.

VAN RIJSBERGEN, C. J., 1979. *Information Retrieval*. Butterworth-Heinemann.