Vesselness for Text Detection in Historical Document Images

Simon Hofmann, Martin Gropp, David Bernecker, Christopher Pollin, Andreas Maier, Vincent Christlein

To cite this version:

S. Hofmann, M. Gropp, D. Bernecker, C. Pollin, A. Maier and V. Christlein, ”Vesselness for text detection in historical document images,” 2016 IEEE International Conference on Image Processing (ICIP), Phoenix, AZ, 2016, pp. 3259-3263.

Submitted on May, 2016

DOI: 10.1109/ICIP.2016.7532962

VESSELNESS FOR TEXT DETECTION IN HISTORICAL DOCUMENT IMAGES

Simon Hofmann[⋆] *Martin Gropp*[⋆] *David Bernecker*[⋆] *Christopher Pollin*[†] *Andreas Maier*[⋆]
Vincent Christlein[⋆]

[⋆] Friedrich-Alexander-Universität Erlangen-Nürnberg,
Pattern Recognition Lab, Martensstraße 3, 91058 Erlangen, Germany
{firstname.lastname}@fau.de
[†] University of Graz, Centre for Information Modelling, Elisabethstraße 59/II, 8010 Graz, Austria
christopher.pollin@uni-graz.at

ABSTRACT

Text detection is typically the first step for any text processing such as hand-written text recognition, layout analysis, line detection, or writer identification. This paper describes a new method to detect text in images, particularly in historical document images. For a robust detection, we propose the use of the vesselness filter as a new preprocessing step for text detection. We show, that this step improves the detection rate significantly. At the locations segmented by this filter, SIFT keypoints are detected which are spatially clustered. Overlapping windows from these clusters are subsequently VLAD encoded and classified in text and non-text. We evaluate this approach on a newly created database, where we achieve an F_1-score of 92%. Additionally, we demonstrate the effectiveness of this method for line segmentation.

Index Terms— text detection, vesselness, historical document analysis, RootSIFT, VLAD

1. INTRODUCTION

With the rise of the digitization of archival material such as books or charters, historical research now often relies on document analysis systems to gain new insights.

The localization of text parts within a document image is an important preprocessing step for further tasks like handwritten text recognition (HTR) or OCR, line segmentation [1], or writer retrieval [2]. Employing text detection limits further processing to relevant areas. This not only reduces the overall computational cost but it can actually be essential for some tasks like writer retrieval, where writers should be recognized by their handwriting and not on the basis of the image background. In contrast to images of modern documents one has to face additional challenges when working on historical documents. The documents are often very heterogeneous and may contain artifacts such as folds, rips or stains. In order to achieve good localization results it is crucial to filter out such artifacts.

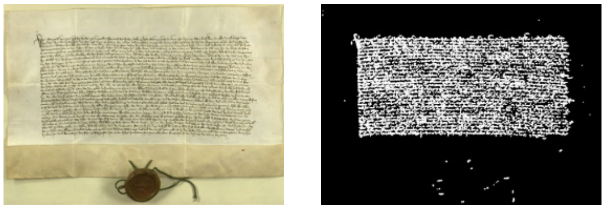

Fig. 1: Input image (left) and corresponding vesselness response (right)

Text detection systems can roughly be separated into two distinct groups. Text detection for printed or handwritten documents on the one hand, and text detection "in the wild", i. e., working on scene images and videos, on the other hand. Text detection in the wild has gained a lot of attention recently [3, 4, 5, 6], while text detection for handwritten documents is typically less prominently embedded in the layout analysis process [7, 8, 9, 10, 11, 12].

Our approach belongs to the latter group, however focusing specifically on the requirements for historical documents, where blocks of typically handwritten text need to be separated from decorations or other non-text elements. For this purpose we propose a new method for text detection. It makes use of a technique to detect *vesselness* [13], a measure initially used for the enhancement of two- and three dimensional images of the human vasculature in medical image processing, which can be reinterpreted as a measure for the probability of a pixel being text. We show that this method is an effective way to filter text-like structures, see Fig. 1, and propose vesselness filtering as an additional preprocessing step for text detection.

From the text candidates found in this step we extract RootSIFT descriptors [14], which are subsequently grouped by the use of DBSCAN [15]. Instead of classifying complete clusters, we tile the clusters into overlapping windows, encode them by means of *vectors of aggregated descriptors* (VLAD) and classify these windows in text or non-text, respectively. This gives a more finegrained text mask.

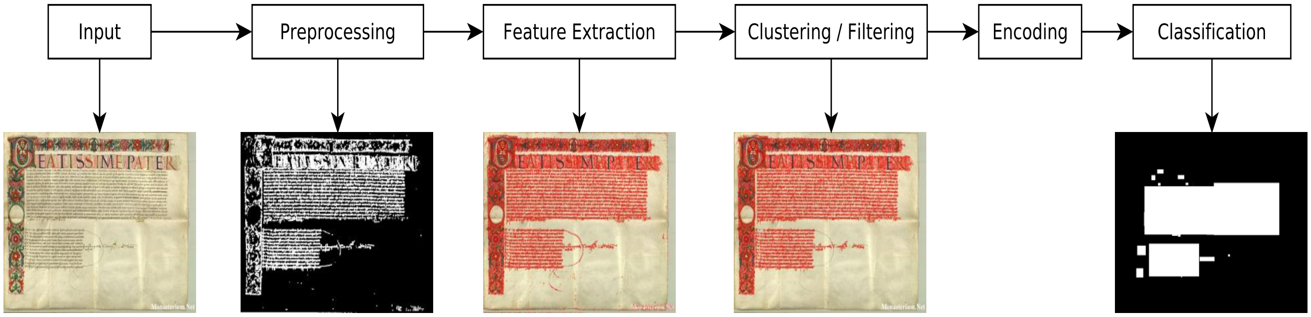

Fig. 2: Detection pipeline

In detail, our contributions are:

- a novel method for text detection in handwritten documents,
- a novel preprocessing step to select candidate text regions using vesselness,
- the illustration of our text detection method as an important preprocessing step for automatic line detection.
- a new dataset of historical document images featuring annotated text areas for evaluation

The rest of this document is structured as follows. Related work is depicted in Sec. 2. Sec. 3 describes the full processing pipeline in detail. Our experiments and results will be discussed in Sec. 4 and concluded in Sec. 5.

2. RELATED WORK

In the layout analysis approach proposed by Wei et al. [10] historical document images are segmented in four classes (periphery, decoration, text, background). Hereby many different feature sets such as local binary pattern (LBP) were used which were reduced by using a feature selection technique which combines a greedy forward selection and a genetic selection to find a very small subset of features.

A layout analysis approach which focuses on text with complex layouts such as side-note texts was proposed by Bukhari et al. [11]. They use different features extracted from the connected components in order to classify different text blocks using a multilayer perceptron (MLP). An MLP has also been used by Baechler and Ingold [7] to classify color and positional features in a multi resolution approach.

Instead of using handcrafted features, Chen et al. [12] segment historical document images in text blocks blocks by using convolutional autoencoders for learning features. The features are then used as input to an SVM. A closely related approach to ours was proposed by Garz et al. [9]. Similarly to our method, they use SIFT descriptors to distinguish between decorative elements and text. Text elements are further grouped into text lines using DBSCAN. In contrast to this approach we are interested in the pure text detection, i. e., the classification between text and non-text. Furthermore, we propose to use vesselness as an initial filtering step and VLAD to get a more reliable classification result.

Typically, these publications use own datasets or publicly available datasets with own annotations. We provide a new publicly available dataset containing medieval handwritten documents with ground truth annotations for the text elements to enable a comparison among different text detection algorithms.

3. METHODOLOGY

The pipeline works as illustrated in Fig. 2. First, we employ the vesselness filter on the grayscale document image, which gives a probability for each pixel being text. We threshold the probability map to achieve an initial text mask. RootSIFT keypoints are extracted from the text mask, and afterwards, keypoints are clustered by their spatial location. Each of the created clusters is then taken as input to a sliding window approach, generating feature descriptors at the respective keypoint locations. These feature descriptors are further aggregated and encoded using VLAD and classified as either text or non-text by a linear classifier. The following sections will describe the different pipeline stages in detail.

3.1. Preprocessing

We propose a novel preprocessing step which significantly improved our pipelines precision. Based on the realization that shape and structure of handwriting is very similar to blood vessels in CTA and MRA images, we implemented a preprocessing step based on Frangis multiscale vessel enhancement filtering [13] to detect candidate regions.

First, the Hessian H is computed from the image at multiple scales as follows:

$$H(f) = \begin{pmatrix} \frac{\partial f}{\partial^2 x} & \frac{\partial f}{\partial x \partial y} \\ \frac{\partial f}{\partial y \partial x} & \frac{\partial f}{\partial^2 y} \end{pmatrix} . \tag{1}$$

From the eigenvalues λ_1, λ_2 of H two measures are extracted: the *dissimilarity measure* (R_b) and the *second order structureness* (S):

$$R_b = \frac{\lambda_1}{\lambda_2} \tag{2}$$

$$S = \sqrt{\lambda_1^2 + \lambda_2^2}\,, \tag{3}$$

where $|\lambda_1| \leq |\lambda_2|$. Based on these two values the filter calculates the *vesselness measure* $\mathcal{V}_i(s)$ that denotes the probability for vessel like structures at location i and scale s.

$$\mathcal{V}_i(s) = \begin{cases} 0 & \text{if } \lambda_2 > 0 \\ \exp(-\frac{R_b^2}{2\beta^2})(1 - \exp(-\frac{S^2}{2c^2}) & \text{otherwise,} \end{cases} \tag{4}$$

where β and c are constants to adjust the filters sensitivity.

Since handwritten characters display a similar tubular structure like vessels, this filter has shown to be very suitable to detect regions of interest for further keypoint extraction. Therefore, we apply a threshold of 0.75 to the probability map and use it to mask regions of interest while filtering out noise. An exemplary filter response for handwritten text is shown in Fig. 3. In comparison to the well known *stroke width transform* (SWT) [3], which computes the stroke width between two parallel edges, the vesselness filter returns a probabilty for each pixel being text. SWT highly depends on the filter parameters to remove false strokes and group strokes with similar width together which makes it very error prone to use.

3.2. Feature Extraction

In order to distinguish parts of text from non-text areas, RootSift [14] descriptors are used. RootSIFT is based upon SIFT [16], additionally normalized with the Hellinger kernel to encounter visual bursts. Visual bursts occur when few large components of the SIFT histogram dominate the similarity computation between two vectors. In practice, each SIFT descriptor is l_1-normalized followed by an element-wise application of the square-root. The descriptors are calculated at keypoints located within the areas masked by the filter result from the previous stage. Subsequently, the descriptors are dimensionality reduced to 64 components using PCA.

3.3. Spatial Clustering

After successful extraction, the generated descriptors are clustered into distinct groups. Detected RootSIFT keypoints show a very dense spatial structure around text (see Fig. 2), which we exploit by performing a density-based clustering (DBSCAN) [15] on keypoint coordinates. Since noise data points are skipped during the clustering process, this clustering step can be seen as an additional filtering step.

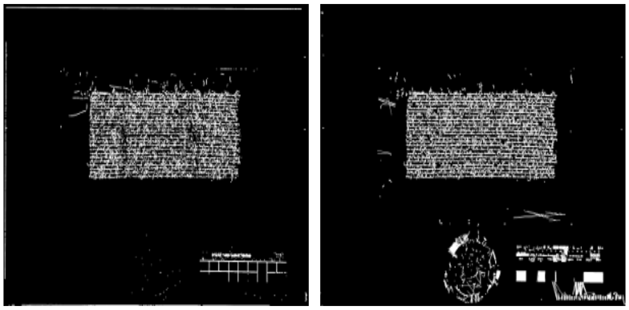

Fig. 3: Vesselness filter response (left) and SWT filter response (right)

3.4. Encoding

The pipelines encoding step applies the basic principle of the bag-of-words (BoW) model in order to represent locally extracted descriptors as a vector of fixed length. As encoding method, we chose to use *vectors of locally aggregated descriptors* (VLAD). These vectors are formed by aggregating the residuals of each local descriptor and its nearest cluster center. VLAD is strongly connected to Fisher vectors and can be seen as a non-probabilistic version of the Fisher Kernel [17]. However, VLAD yields a more compact representation than Fisher vectors. With improvements like whitening [18] or intra-normalization [19] VLAD achieves state of the art performance on several benchmark datasets. Note that we employ both of these techniques.

Let $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_T\}$ denote the T local image descriptors $\mathbf{x}_t \in \mathbb{R}^N$. A dictionary $\mathbf{D} = \{\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K\}$, consisting of K clusters $\boldsymbol{\mu} \in \mathbb{R}^N$, is computed using k-means from a random subset of local image descriptors. Each local descriptor is then assigned to its nearest cluster center. For each cluster k the differences between its cluster center and the assigned local descriptors are aggregated [17]:

$$\mathbf{v}_k = \sum_{\mathbf{x}_t:\ \mathrm{NN}(\mathbf{x}_t)=\boldsymbol{\mu}_k} (\mathbf{x}_t - \boldsymbol{\mu}_k)\,, \tag{5}$$

where $\mathrm{NN}_k(x_t)$ denotes the nearest cluster center of $\mathbf{x}_t$ in the dictionary $\mathbf{D}$. The final global vector follows as: $\mathbf{v} := \left(\mathbf{v}_1^\top, \ldots, \mathbf{v}_K^\top\right)^\top$. This global vector is further postprocessed. First we employ a normalization step to counter visual bursts. Arandjelović and Zisserman [19] proposed the use of *intra-normalization* where each $\mathbf{v}_k$ is l_2 normalized individually before the final concatenation.

Additionally, we decorrelate the global vector by performing a PCA-whitening. This helps to dampen co-occurrences. Hereby, we follow the approach of Jégou and Ondřej [18]. Instead of using a single dictionary, multiple dictionaries (in our

Method	Precision	Recall	F_1
RootSIFT + VLAD	0.82	**0.95**	0.88
SWT + RootSIFT + VLAD	0.75	0.92	0.83
Vesselness + RootSIFT + VQ1k	0.78	0.88	0.83
Vesselness + RootSIFT + VQ10k	0.8	0.79	0.79
Vesselness + RootSIFT + VLAD	**0.94**	0.91	**0.92**

Table 1: Text detection performance on MomDB.

case: three) are used for the encoding step, i. e., for each dictionary an individual VLAD representation is computed. Subsequently, these representations are concatenated and decorrelated. This has been shown to be very beneficial for image retrieval [18, 20].

3.5. Classification

Encoded features are classified by a linear SVM. For our pipeline, we used the implementation provided by vlfeat [21] which includes the PEGASOS algorithm proposed by Shalev-Shwartz et al. [22], an algorithm especially suitable for training with many samples.

4. EVALUATION

We evaluate the proposed method for text detection on two different datasets. First, text is detected on a newly created dataset containing medieval document images. Second, the improvement of a line segmentation system is shown as an example for the importance of a robust text detection system.

4.1. Dataset

Currently, there is no publicly available dataset that has annotations for all text occurences. Thus, we propose the use of the new dataset MomDB. This very heterogeneous dataset consists of 249 images from Stiftsarchiv Geras, documenting over 800 years of Austrian monasterial history between 1188 and 1992. It includes both handwritten and printed pages, all with annotated text areas. The images in this set were provided through the collaborative online archive *Monasterium*[1].

4.2. Text Detection

Tab. 1 shows the average pixel wise precision and recall for MomDB. While using the proposed vesselness filter as a preprocessing step slightly impairs the recall, as expected, it considerably improves precision. This results in an increased overall F_1-score. Using SWT instead the vesselness filter even worsens the recognition rate.

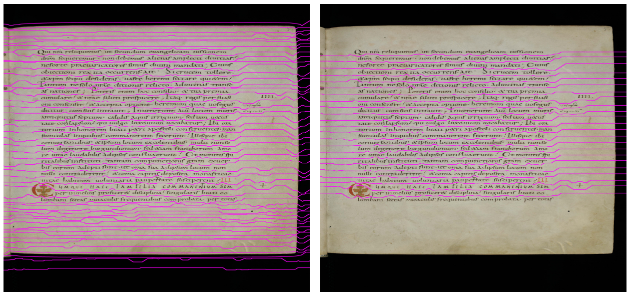

Fig. 4: Line segmentation without text detection (left) and with prior text detection (right).

Another interesting result is the comparision between VLAD and traditional *vector quantization* (VQ). Vector quantization denotes the normalized histogram of visual words. For the clustering step we used k-means with 1000 (VQ1k) and 10000 (VQ10k) cluster centers (for VLAD 64). Tab. 1 shows that VLAD drastically improves the recognition performance in comparison to vector quantization.

4.3. Line Segmentation

The detection of text areas serves as a preprocessing step to other text processing algorithms such as line segmentation. Thus, we also evaluated the improvement of a typical line segmentation system when the text areas are pre-segmented. We used the Saint Gall database SGDB [23] and employed our pipeline to generate masks of text areas in order to process only regions of interest. The test data consists of 60 single page images with additional ground truth data for evaluation.

We followed the evaluation protocol of the ICDAR handwriting segmentation contest [24] with a minimum overlap of 75%. When using the text detection method prior to the line segmentation method [25], we achieve an F_1-score of 93%. In contrast, we only achieve an accuracy of 54% when the line segmentation algorithm is used directly on the data. A qualitative example can be viewed at Fig. 4, which shows that the line segmentation is only performed on the actual text area.

5. CONCLUSION

In this work, we have presented a new framework for text detection in historical handwritten document images. First, we introduce a new preprocessing strategy for filtering text using the vesselness filter. We showed that this greatly improves the detection results. Second, we propose VLAD encoded RootSIFT descriptors to form a higher dimensional feature vector which simplifies the classification task. To evaluate our system, we created a new dataset containing medieval document images and marked ground truth text areas. In future, we would like to combine vesselness with other strategies like SWT or Stroke Support Pixels [26].

[1] http://monasterium.net

6. REFERENCES

[1] O. Surinta, M. Holtkamp, F. Karabaa, J.-P. Van Oosten, L. Schomaker, and M. Wiering, "A* Path Planning for Line Segmentation of Handwritten Documents," in *ICFHR*, Sep 2014, pp. 175–180. 1

[2] V. Christlein, D. Bernecker, and E. Angelopoulou, "Writer identification using vlad encoded contour-zernike moments," in *ICDAR*, Aug 2015, pp. 906–910. 1

[3] B. Epshtein, E. Ofek, and Y. Wexler, "Detecting text in natural scenes with stroke width transform," in *CVPR*, Jun 2010, pp. 2963–2970. 1, 3

[4] M. Jaderberg, A. Vedaldi, and A. Zisserman, "Deep Features for Text Spotting," in *ECCV*, vol. 8692, pp. 512–528. Sep 2014. 1

[5] X.-C. Yin, X. Yin, K. Huang, and H.-W. Hao, "Robust Text Detection in Natural Scene Images," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 36, no. 5, pp. 970–983, May 2014. 1

[6] M. Bušta, L. Neumann, and J. Matas, "FASText: Efficient Unconstrained Scene Text Detector," in *ICCV*, Jun 2015. 1

[7] M. Baechler and R. Ingold, "Multi Resolution Layout Analysis of Medieval Manuscripts Using Dynamic MLP," in *ICDAR*, Sep 2011, pp. 1185–1189. 1, 2

[8] A. Garz, M. Diem, and R. Sablatnig, "Detecting Text Areas and Decorative Elements in Ancient Manuscripts," in *ICFHR*, Nov 2010, pp. 176–181. 1

[9] A. Garz, R. Sablatnig, and M. Diem, "Layout Analysis for Historical Manuscripts Using SIFT Features," in *ICDAR*, Sep 2011, pp. 508–512. 1, 2

[10] H. Wei, K. Chen, R. Ingold, and M. Liwicki, "Hybrid Feature Selection for Historical Document Layout Analysis," in *ICFHR*, Sep 2014, pp. 87–92. 1, 2

[11] S. S. Bukhari, T. M. Breuel, A. Asi, and J. El-Sana, "Layout analysis for arabic historical document images using machine learning," in *ICFHR*, Sep 2012, pp. 639–644. 1, 2

[12] K. Chen, M. Seuret, M. Liwicki, J. Hennebert, and R. Ingold, "Page Segmentation of Historical Document Images with Convolutional Autoencoders," in *ICDAR*, Sep 2015, pp. 1011–1015. 1, 2

[13] A. F. Frangi, W. J. Niessen, K. L. Vincken, and M. A. Viergever, "Multiscale Vessel Enhancement Filtering," in *Medical Image Computing and Computer-Assisted Intervention — MICCAI'98*, vol. 1496 of *Lecture Notes in Computer Science*, pp. 130–137. 1998. 1, 2

[14] R. Arandjelović and A. Zisserman, "Three things everyone should know to improve object retrieval," in *CVPR*, 2012, pp. 2911–2918. 1, 3

[15] Martin E., H.-P. Kriegel, J. Sander, and X. Xu, "A Density-based Algorithm for Discovering Clusters in Large Spatial Databases with Noise," in *Knowledge Discovery and Data Mining, 2nd International Conference on*, 1996, pp. 226–231. 1, 3

[16] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004. 3

[17] H. Jégou, F. Perronnin, M. Douze, J. Sánchez, P. Pérez, and C. Schmid, "Aggregating Local Image Descriptors into Compact Codes," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 34, no. 9, pp. 1704–1716, 2012. 3

[18] H. Jégou and O. Chum, "Negative Evidences and Co-occurences in Image Retrieval: The Benefit of PCA and Whitening," in *ECCV*, pp. 774–787. 2012. 3, 4

[19] R. Arandjelović and A. Zisserman, "All about VLAD," in *CVPR*, 2013, pp. 1578–1585. 3

[20] E. Spyromitros-Xioufis, S. Papadopoulos, I. Kompatsiaris, G. Tsoumakas, and I. Vlahavas, "A Comprehensive Study Over VLAD and Product Quantizationin Large-Scale Image Retrieval," *Multimedia, IEEE Transactions on*, vol. 16, no. 6, pp. 1713–1728, 2014. 4

[21] A. Vedaldi and B. Fulkerson, "VLFeat: An Open and Portable Library of Computer Vision Algorithms," `http://www.vlfeat.org/`, 2008. 4

[22] S. Shalev-Shwartz, Y. Singer, N. Srebro, and A. Cotter, "Pegasos: Primal Estimated Sub-gradient Solver for SVM," *Mathematical programming*, vol. 127, no. 1, pp. 3–30, 2011. 4

[23] A. Fischer, V. Frinken, A. Fornés, and H. Bunke, "Transcription alignment of latin manuscripts using hidden markov models," in *Historical Document Imaging and Processing, Workshop on*. 2011, HIP '11, pp. 29–36, ACM. 4

[24] N. Stamatopoulos, B. Gatos, G. Louloudis, U. Pal, and A. Alaei, "ICDAR 2013 Handwriting Segmentation Contest," in *ICDAR*, aug 2013, pp. 1402–1406. 4

[25] N. Arvanitopoulos and S. Susstrunk, "Seam carving for text line extraction on color and grayscale historical manuscripts," in *ICFHR*, Sept 2014, pp. 726–731. 4

[26] L. Neumann and J. Matas, "Efficient scene text localization and recognition with local character refinement," in *ICDAR*, Aug 2015, pp. 746–750. 4