Page Segmentation for Historical Handwritten Document Images Using Color and Texture Features

Kai Chen*, Hao Wei*, Jean Hennebert*†, Rolf Ingold*, and Marcus Liwicki*‡
*DIVA (Document, Image and Voice Analysis) research group
Department of Informatics, University of Fribourg, Switzerland
Email: {firstname.lastname}@unifr.ch
†University of Applied Sciences, HES-SO//FR, Bd. de Pérolles 80, 1705 Fribourg, Switzerland
‡DFKI - German Research Center for Artificial Intelligence

***Abstract*—In this paper we present a physical structure detection method for historical handwritten document images. We considered layout analysis as a pixel labeling problem. By classifying each pixel as either *periphery*, *background*, *text block*, or *decoration*, we achieve high quality segmentation without any assumption of specific topologies and shapes. Various color and texture features such as color variance, smoothness, Laplacian, Local Binary Patterns, and Gabor Dominant Orientation Histogram are used for classification. Some of these features have so far not got many attentions for document image layout analysis. By applying an Improved Fast Correlation-Based Filter feature selection algorithm, the redundant and irrelevant features are removed. Finally, the segmentation results are refined by a smoothing post-processing procedure. The proposed method is demonstrated by experiments conducted on three different historical handwritten document image datasets. Experiments show the benefit of combining various color and texture features for classification. The results also show the advantage of using a feature selection method to choose optimal feature subset. By applying the proposed method we achieve superior accuracy compared with earlier work on several datasets, e.g., we achieved 93% accuracy compared with 91% of the previous method on the *Parzival* dataset which contains about 100 million pixels.**

***Keywords*-page segmentation; historical document; layout analysis; feature selection;**

I. Introduction

In recent years, a large number of historical documents have been digitized and made available to the public. With the increasing availability of computers and text-based software, the analysis of such documents is leveraged to higher dimensions leading to novel interests in digital humanities research. Layout analysis is considered as an important initial step for content recognition. It aims at splitting a page image into regions of interest and distinguishing text blocks from the regions. Due to the complex layout, degradation of the page, and different writing styles, layout analysis on the historical documents is a challenging task that has received a significant amount of attention. Our goal is to develop a generic, flexible, and robust segmentation method to delimit text blocks, text lines, and eventually isolated words[1].

In this paper, we propose a layout structure segmentation algorithm which is applicable to color historical handwritten document images with various layout. Most of the state-of-the-art methods of page segmentation on historical documents are based on the connected components aggregation on the binary image [1], i.e., a binarization pre-processing is needed. Unlike these methods, our method is applicable on color images, i.e., the proposed method is applied directly on the color images without any binarization pre-processing. We consider the segmentation as a pixel classification problem, where each pixel is represented by a vector containing features based on the coordinates, color, and texture information gathered from the neighbourhood of the pixel. By training the classifier with these features, we classify each pixel into one of the four classes: *periphery*, *background*, *text block*, and *decoration*. Extending our earlier work [2], [14], for the segmentation we use coordinates, color, and texture features, i.e., color variance, smoothness, Laplacian, Local Binary Patterns and Gabor Dominant Orientation Histogram which have not gotten many attention for historical document image layout analysis. An improved Fast Correlation-Based Filter feature selection algorithm is also applied in order to reduce the feature size. Our main idea is to use many different kinds of features and let the feature selection method selecting the optimal subset of features. Finally, we apply a smoothing post-processing procedure to smooth out noisy classification. Our experimental results show that the proposed method is superior to the previous method described in [14].

In summary, the main contributions of this work consist of: (1) Besides using coordinates and the maximum and minimum color features as described in [14], we also use texture features (i.e., Local Binary Pattern and Gabor Dominant Orientation Histogram) of each pixel's neighborhood for classification. Other color features such as: variance, smoothness, and Laplacian are also combined in order to

[1]HisDoc: Historical Document Analysis, Recognition, and Retrieval. https://diuf.unifr.ch/main/hisdoc/hisdoc2

2167-6445/14 $31.00 © 2014 IEEE
DOI 10.1109/ICFHR.2014.88

improve the performance. (2) We improve a state-of-the-art feature selection method to select the optimal feature subset for different datasets. (3) We introduce a post-processing approach to refine the results. Experiments are performed on three different historical handwritten document image datasets [5]: *Parzival*[2], *George Washington*[3], and *Saint Gall*[4]. The experimental results show that the proposed method is effective and robust to changes of writing style, page layout, and noise on the images. We conclude experimentally that coordinates, color, and texture are crucial information for page segmentation on historical manuscript images.

The remainder of this paper is organized as follows. Section II gives an overview of some related works in layout analysis for historical document images. Section III describes the proposed page segmentation method. Section IV reports on our experimental results and Section V presents conclusions and future works.

II. Related Works

In this section we discuss several state-of-the-art works dealing with layout analysis of historical documents.

AGORA [12] uses two maps to segment historical document images: a shape map that focuses on connected components and a background map which provides information about white areas corresponding to block separations in the page. Then it uses simultaneously the information provided by the two maps for segmentation. After segmentation a list of blocks are created. Users are able to label, merge, and remove them. DEBORA [9] aims at improving the accessibility of rare sixteen century books through the Internet. It uses image analysis to extract documents metadata. Compression is realized by analysis of their content. Their segmentation task includes the segmentation of text from non text, segmentation of the main text body from margins, and physical layout segmentation. Grana et al. [8] present a system for automatic segmentation, annotation, and image retrieval based on content, focused on illuminated manuscripts and in particular the Borso D'Este Holy Bible. Documents are mainly divided into three parts: background, text, and decorations. They use some texture analysis techniques based on circular statistics to segment handwritten text and illustrations. They also propose a user interface for browsing the pages through visual similarity. In the Historical Document Layout Analysis Competition (ICDAR 2011) [1], four layout analysis methods for printed historical document images were evaluated and compared with a state-of-the-art commercial software. The results indicate that there is a convergence to a certain methodology with some variations in the approach, i.e., connected components aggregation on the binary images. However, it is also clear that there is still a considerable need to develop robust methods for layout analysis on historical documents.

[2]http://www.parzival.unibe.ch

[3]http://memory.loc.gov/ammem/gwhtml/gwseries2.html

[4]http://www.e-codices.unifr.ch

III. System description

In this work, we consider page segmentation as a pixel classification problem. Due to the large size of the images (at least 1500 $\times$ 2000 pixels for each image), layout analysis is time consuming. As our segmentation method will be used for ground-truth generation and it will be embedded into a GUI, the algorithm will be used online and has to be computationally efficient. For this reason, we based our work on the pyramidal approach of [2]. At the first level, we scale each image to a smaller size with the scale factor $\alpha < 1.0$. Then the scaled image is segmented into four parts, i.e., *out of page*, *background*, *text block*, and *decoration*. At the second level, the image has the double resolution of the first level in order to perform the more precise tasks such as text line segmentations. Our contribution is mainly focused on the first level.

A. Feature Extraction

Feature extraction is an important part of the classification task. In order to build a flexible and robust page segmentation system for different historical documents, we investigate various features for classification. In the proposed method, each pixel $p_{x,y}$ is represented by a d-dimensional real-valued feature vector which is computed from its neighborhood. For a given pixel $p_{x,y}$, its neighbors $N(p_{x,y}, n)$ are the pixels in a $n \times n$ window, $N(p_{x,y}, n)$ is defined as: $N(p_{x,y}, n) = \{p_{x',y'} | x' \in \{x-d, x-d+1, \cdots, x+d-1, x+d\} \wedge y' \in \{y-d, y-d+1, \cdots, y+d-1, y+d\} \wedge x' \neq x \wedge y' \neq y \wedge d = (n-1)/2\}$. The feature vector is composed by concatenating the features in three categories: coordinates, color, and texture. For a given pixel $p_{x,y}$, these features are defined as follows.

1) Coordinate: All the images in the same dataset are normalized to the same size with a scale factor α, then for each pixel, its x and y coordinates are used as features.

2) Color: Since we directly work on color images, color is considered as an important information for classification. The features used in the previous work are: primary colors value of r, g, and b, sum of neighborhood, maximum and minimum of neighbor pixels, and sum of pixels in the column of the whole document. The details of these features are explained in [14]. Since our objective is to create a generic page segmentation system for various historical documents, therefore in order to get more color information around the pixel, we extend these color features. Several new color features are employed in this work. These features are:

- Mean value of neighborhood primary color, e.g., the mean value of r component is given as $M(p_{x,y}, n)^r = \frac{S(p_{x,y}, n)^r}{n \times n}$, where $S(p_{x,y}, n)^r$ returns the sum of r component of $N(p_{x,y}, n)$.

- Variance of neighborhood primary, e.g., the variance of r component is give as: $V(p_{x,y},n)^r = \frac{\sum_{i=-d}^{d}\sum_{j=-d}^{d}(z^r_{x+i,y+j}-M(p_{x,y})^r)^2}{n\times n}$, $d=(n-1)/2$.
- Color smoothness [6] is a transformation of the variance. It is defined as: $SMO(p_{x,y},n)=\frac{1}{1+V(p_{x,y},n)}$.
- Horizontal mean, variance, and smoothness of neighborhood primary color. We only use the neighbors on the horizontal direction of $p_{x,y}$ to compute the values of the mean and the variance.
- Laplacian [13] is the sum of second partial derivative of $z_{x,y}$ on x and y, where $z_{x,y}$ is the pixel value function on position x and y. The Laplacian is used to extract the information about the speed of color variation on the x and y direction. It is defined as:

$$\bigtriangledown^2 z(x,y)=\frac{\partial^2}{\partial x^2}z(x,y)+\frac{\partial^2}{\partial y^2}z(x,y) \quad (1)$$

3) Texture: In order to extract the texture features, we convert the color images into gray levels. Then we investigate two type of texture features.

Local binary patterns (LBP) is a feature which has been widely used for texture analysis [10]. LBP is based on signs of differences in a circular neighboring pixels. It is defined as:

$$LBP_{P,R}(x,y)=\sum_{p=0}^{P-1}s(g_p-g_c)2^p$$
$$s(x)=\begin{cases}1 & x\geq 1\\ 0 & x<0\end{cases} \quad (2)$$

For a given pixel $p_{x,y}$, P is defined as the number of neighbors; R is defined as the radius of a circle centered on $p_{x,y}$; g_c is the gray level value of the central pixel at coordinates x, y; g_p is the gray level value of the p-th point on the circular neighborhood. In order to capture the local structure information (e.g., edges, lines, and spots) of pixel $p_{x,y}$ in its neighbors $N(p_{x,y},n)$, we compute the LBP histogram based on the Rotation Invariant Unifrom Pattern (LBP^{riu2}) [11].

The output of LBP^{riu2} can be used to measure the spatial pattern, but it discards contrast. In order to overcome this drawback, we compute the local variance as described in [11]. The local variance of pixel $p_{x,y}$ is given as:

$$VAR^{riu2}_{P,R}=\frac{1}{P}\sum_{p=0}^{P-1}s(g_p-\mu), \quad \text{where} \quad \mu=\frac{1}{P}\sum_{p=0}^{P-1}g_p \quad (3)$$

The number of bins of the histogram is $P+2$. In this work, we choose $P=24$, $R=3$.

Gabor Dominant Orientation Histogram. Gabor filter has been widely applied in classification tasks, e.g., texture analysis, face recognition, character recognition, and moving object tracking. It extracts orientation-dependent frequency content by performing a spatial frequency analysis. As described in [7], Gabor filters localize direction spatial frequency at orientation θ, i.e., the output of Gabor filter $h_\theta(x,y)$ to an image responds maximally at those edges of the orientation θ. In contrast to [4], for a given pixel $p_{x,y}$, instead of using its dominant orientation as the feature, in this work we compute the dominant orientation histogram on its neighborhood $N(p_{x,y},n)$. The histogram is computed as follows.

First, for each pixel $p_{x',y'}$ in $N(p_{x',y'},n)$, we compute the sum of the convolution of a set of Gabor filters $h_\theta(x',y')$ on different orientations. We define $I_\theta(x',y')$ as the sum of convolution of Gabor filter $h_\theta(x',y')$ on an image $u(x',y')$ at the orientation θ, where $I_\theta(x',y')$ is defined as:

$$I_\theta(x,y)=\sum_{i=x-\frac{n-1}{2}}^{x+\frac{n-1}{2}}\sum_{j=y-\frac{n-1}{2}}^{y+\frac{n-1}{2}}u(i,j)e^{\frac{-(i-x)^2-(j-y)^2}{2\sigma^2}} e^{j\lambda(cos\theta(i-x)+sin\theta(j-y))} \quad (4)$$

λ is the wavelength of the Gabor filter and σ is the standard derivation of Gaussian along the x and y direction. The orientation angles of this set of Gabor filters are: $\{\theta_k|\theta_k=k\times\frac{2\pi}{K-1}\}$, where $k=0,\cdots,(K-1)$. If we obtain a maximum output $I_{\theta_k}(x',y')$, then θ_k is considered as the dominant orientation angle of the pixel at position (x',y'). Then we accumulate the occurrence of θ_k in $N(p_{x,y},n)$ to generate a histogram, where the number of bins is $|\theta_k|$. In our system, we choose $K=10$, $\lambda=1$, and $\sigma=2$. For details of how to select λ and σ, we refer to [7].

B. Feature Selection

Due to the high dimensionality of features and large amount of pixels in the images, a feature selection method is applied to reduce the computational time. The objective of the feature selection is to reduce the dimensionality of the feature space without decreasing the accuracy by removing irrelevant and redundant features. In this work, we improved the Fast Correlation-Based Filter (FCBF) algorithm [15] for optimal feature subset selection. FCBF generates a subset of features S', such that all the features in S' are highly correlated to the classes but they are uncorrelated to any of the other features. The correlation measure is based on the information-theoretical concept of *entropy*, i.e., a measure of the uncertainty of a random variable. In particular, a threshold δ is chosen to measure the *predominant correlation* for a feature to a class. High *predominant correlation* value features are selected as the optimal feature subset. For the definition of *predominant correlation* and details of FCBF algorithm, we refer to [15].

Choosing δ is crucial for this algorithm. If δ is too large some important features may be removed. On the other hand, if δ is too small, some redundant features may not be removed. Furthermore, since the characteristics of various datasets are different, δ is dependent on the dataset. The

optimal δ value can be discovered by exhaustive experimental search. We propose here an alternative procedure to efficiently compute δ. We set δ to a small value (e.g., 0.01) at the beginning. In each iteration, we increase the value of δ, such that $\delta = \delta + \gamma$, where γ is also a small value (e.g., 0.01), then we apply FCBF algorithm on the full feature set S to generate the optimal feature subset S_{best} iteratively until the size of S_{best} become stable, i.e., the size of S_{best} did not change. We expect that the optimal *predominant correlation* value δ is different from one dataset to another as the optimal feature subset is dependent on the quality and characteristic of documents.

C. Classification

The focus of our work is in the proposition of efficient features and their selection as described above. Therefore, we used for the classification part default settings of a state-of-the-art SVM implementation [3].

D. Post-Processing

We observe that the pixels on the border of the text are usually degraded. By observing the page segmentation results on the images in the validation set, we find some errors of classification consistently appears on these pixels. In order to correct this type of errors, we proposed a post-processing method as follows.

After page segmentation, pixels are classified into four classes: *periphery*, *background*, *text*, and *decoration*. For a non *text* pixel p, we compute the number of *text* pixels NT in its neighborhood $N(p, n')$. If $\frac{NT}{n' \times n'} \geq T$, then p is labelled to *text*, where $n' = 3$ and $T = 0.6$ are used in our system. We applied the same approach on the non *periphery* pixels.

IV. Experiments

We now present the experimental results achieved with the system described above. We demonstrate the impact of using different features for page segmentation on the three different historical document image datasets. These datasets are the same as the ones which have been used for the evaluation in the earlier work [14].

A. Datasets

The datasets are of very different nature. The *George Washington* dataset consists of 20 pages written in English with ink on paper and the images are in gray levels. The image size is 2200×3400 pixels. The other two datasets, i.e., *Parzival* [5] and *Saint Gall* consist of images of manuscripts written with ink on parchment and the images are in color, while the former suffers from many degradations. The *Parzival* dataset consists of 47 pages written by three writers. These pages were taken from a medieval German manuscript from the 13th century that contains the epic poem Parzival by Wolfram von Eschenbach. The image size is 2000×3008 pixels. The *Saint Gall* data set consists of 60 manuscript pages from a medieval manuscript written in Latin. The image size is 1664×2496 pixels.

As described in Section III, in our system, images are scaled down with a scale factor α. In the experiments, we use the same scale factor value as given in [14], i.e., $\alpha = \frac{1}{16}$. Figure 1 gives some example pages with their layout ground-truth. Table I gives the details of the training set TR, test set TE, and validation set VA.

(a) (b) (c)

(d) (e) (f)

Figure 1: Page 1a and its ground-truth 1d from the *George Washington* dataset. Page 1b and its ground-truth 1e from the *Saint Gall* dataset. Pages 1c and its ground-truth 1f from the *Parzival* dataset.

Table I: Training, test, and validation sets details.

	page size(pixels)	$\|TR\|$	$\|TE\|$	$\|VA\|$
George Washington	138×213	10	5	4
Saint Gall	104×156	20	30	10
Parzival	138×213	24	14	2

B. Setup

In order to evaluate the impact of different features on the page segmentation, we use various combination of features as described in Section III-A. The features are defined as: F_1, coordinates feature; F_2, color feature; F_3, texture feature. For reporting the results we use the same evaluation metric as used in [14], i.e., the classification accuracy on the pixel level. Precision P and Recall R on the pixel level are also used for the evaluation.

C. Optimal window size estimation

In order to extract distinctive features of the pixel p in its neighbors $N(p, n)$, we estimate the optimal window size n by evaluating the performance with varying numbers of n on the validation set on each dataset. The feature vector is composed with the concatenation of F_1, F_2, and F_3. Table III shows the error rate (%) on the different window size n. We note that the optimal window size n is 7, 11,

Table II: Page segmentation methods comparison. Features are defined as: F_1: coordinates. F_2: color. F_3: texture.

	George Washington		Parzival		Saint Gall	
	Feature size	accuracy (%)	Feature size	accuracy (%)	Feature size	accuracy (%)
Wei et al. SVM	87	88.7	87	90.7	87	96.2
Wei et al. MLP	87	90.4	87	90.1	87	97.5
Wei et al. GMM	87	88.7	87	90.7	87	96.2
Wei et al. Combine	87	89.1	87	91.1	87	97.1
Proposed method with various setups						
F_2	84	88.8	160	92.6	122	97.5
F_2 (post-processing)	84	89.3	160	92.5	122	97.7
$F_1 + F_2$	86	89.6	162	92.8	124	97.8
$F_1 + F_2$ (post-processing)	86	90.1	162	92.8	124	97.9
F_3	38	90.3	38	87.5	38	95.1
F_3 (post-processing)	38	90.4	38	87.8	38	95.2
$F_1 + F_3$	40	90.8	40	88.5	40	95.6
$F_1 + F_3$ (post-processing)	40	91.0	40	88.8	40	95.7
$F_1 + F_2 + F_3$	124	91.5	200	93.0	162	97.8
$F_1 + F_2 + F_3$ (post-processing)	124	91.6	200	93.1	162	**97.9**
$F_1 + F_2 + F_3$ (feature selection)	90	91.7	139	93.2	123	97.8
$F_1 + F_2 + F_3$ (feature selection and post-processing)	90	**91.9**	139	**93.2**	123	**97.9**

and 15 for the *George Washington*, *Saint Gall*, and *Parzival* dataset respectively. These values are used for the page segmentation on the test sets.

Table III: Classification error rate (%) on varying window sizes.

window size	*George Washington*	*Saint Gall*	*Parzival*
3×3	8.3	3.2	8.3
5×5	7.2	2.7	6.6
7×7	**6.4**	3.0	6.1
11×11	7.1	**2.6**	5.1
15×15	7.3	2.9	**4.4**
19×19	7.2	2.7	4.5

D. Experimental Results and Discussion

In this section, we present the experiment results and give discussions. Figure 2 depicts the decrease of optimal feature subset size on the *predominant correlation* value δ with the feature selection algorithm described in Section III-B. In this experiment, the full feature set contains all the features as described in the Section III-A, the window size is set to 10 for the three datasets. We note that δ is dependent on datasets. Due to the different quality and characteristic of the various datasets, it takes six iterations to find the optimal feature subset on the *Parzival* dataset; three iterations on the *George Washington* and *Saint Gall* datasets.

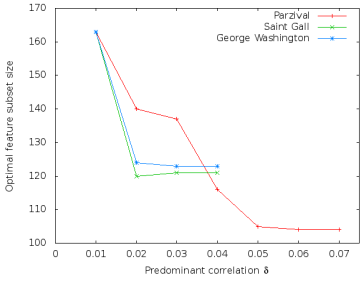


Figure 2: Feature selection algorithm results.

In the earlier work, Wei et al. [14] used only coordinates and color as features. They used three classifiers, i.e., SVM, MLP, and GMM. Finally, they combined the three classifiers to determine the class label, i.e., the final class label is determined by the majority voting on the outputs of the three classifiers. Table II compares the average classification accuracy of the proposed method in different setups with the state-of-the-art methods as described in [14].

On the experimental results, we note that pixels coordinates are important information for layout analysis. The performance is improved when we combine the coordinates with color or texture features.

Texture features are more important for the gray levels images than color images. It is seen that when texture features are combined with coordinates information, the proposed method outperforms all the previous methods on the *George Washington* dataset[5]. Moreover, the texture features size are much smaller than the features size of the previous method as shown in Table II, i.e., $|F1 + F3| = 40$ compared with 87 of the previous method.

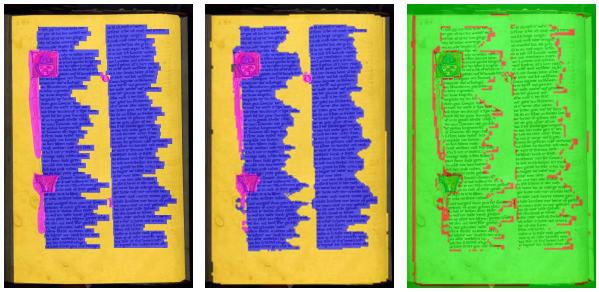
We extend the color features used in the previous work by adding color features, i.e., mean, variance, smoothness, and Laplacian as described in Section III-A2. By using the extended color features, we achieved better results than all the previous methods on the *Parzival* and *Saint Gall* datasets. The result on the *George Washington* is comparable to the previous method of the best performance (i.e., using MLP) and is better than the other methods of the previous work. Note that the method using MLP has the drawback that it can not detect the decoration on the *Parzival* dataset (i.e., $P = 0\%$), but the proposed method achieves $P = 90.6\%$.

From Table II, it is observed that we achieve the best performance when combining the coordinates, color, and texture features. By applying the improved FCBF feature selection method, the feature size is significantly reduced, i.e.,

[5]Note that the *George Washington* dataset contains only gray level images, the other two datasets contain only color images.

27%, 30%, and 24% features are removed for the *George Washington*, *Parzival*, and *Saint Gall* datasets respectively. Moreover, the feature selection algorithm does not degrade the performance on the *Saint Gall* dataset and even improves a little the performance on the *George Washington* and *Parzival* dataset.

Figure 3 shows the page segmentation results of a test image from the *Parzival* dataset with the proposed method, i.e., using all the features, then applying the feature selection algorithm, finally applying the post-processing method.



(a) ground-truth (b) segmentation result (c) errors

Figure 3: Page segmentation results of image in Fig. 1c. Fig. 3c gives the page segmentation results of correctly classified pixels (in Green) and misclassified pixels (in Red).

V. Conclusion

In this paper, we proposed a page segmentation method for the historical handwritten document images. The method has been evaluated on three historical manuscript datasets of diverse nature. The experiments demonstrate the effectiveness and robustness of the method. Based on the coordinates, color, and texture features, we achieved superior performance compared with previous work [14]. Our results show that it is possible to achieve high performance of page segmentation by combining coordinates, color, and texture features. In addition, feature selection algorithms can help to reduce the feature size without degrading the classification accuracy. With more significant features, sophisticated feature selection algorithms, and classifiers currently developed by the computer vision and machine learning researchers, we believe that the approaches pursued here might achieve better performance.

Acknowledgment

We would like to thank Michael Stolz for providing us with the Parzival data set. The project is part of HisDoc and HisDoc 2.0 funded by the SNF via the grant numbers CRSI22_125220 and 205120_150173.

References

[1] A. Antonacopoulos, C. Clausner, C. Papadopoulos and S. Pletschacher, *Historical Document Layout Analysis Competition*, International Conference on Document Analysis and Recognition, 2011, pp. 1516-1520.

[2] M. Baechler and R. Ingold, *Multi Resolution Layout Analysis of Medieval Manuscripts Using Dynamic MLP*, Proceedings of The Eleventh International Conference on Document Analysis and Recognition, 2011, pp. 1185-1189.

[3] C.-C. Chang and C.-J. Lin, *LIBSVM : a library for support vector machines.* ACM Transactions on Intelligent Systems and Technology, 2011, vol. 2, issue 3, pp. 2:27:1–27:27.

[4] K. Chen, H. Wei, M. Liwicki, J. Hennebert, and R. Ingold, *Robust Text Line Segmentation for Historical Manuscript Images using Color and Texture.* 22nd International Conference on Pattern Recognition, 2014, accepted.

[5] A. Fischer, A. Keller, V. Frinken, and H. Bunke, *Lexicon-Free Handwritten Word Spotting Using Character HMMs*, Pattern Recognition Letters, vol. 33(7), 2012, pp. 934-942.

[6] M. Hangarge, K.C. Santosh, S. Doddamani, and R. Pardeshi, *Statistical Texture Features based Handwritten and Printed Text Classification in South Indian Documents*, CoRR, abs/1303.3087, 2013.

[7] P. Hu, Y. Zhao, Z. Yang, and J. Wang, *Recognition of gray character using gabor filters*, Proceedings of the Fifth International Conference on Information Fusion, vol. 1, 2002, pp. 419-424.

[8] C. Grana, D. Borghesani, S. Calderara, and R. Cucchiara, *Inside the Bible: segmentation, annotation and retrieval for a new browsing experience*, ACM international conference on Multimedia Information Retrieval, 2008, pp. 379-386.

[9] F. Le Bourgeois and H. Emptoz, *DEBORA: Digital AccEss to BOoks of the RenAissance*, International Journal of Document Analysis and Recognition (IJDAR), vol. 9, issue 2-4, 2007, pp. 193-221.

[10] T. Ojala, M. Pietikäinen, and D. Harwood, *A Comparative Study of Texture Measures with Classification Based on Feature Distributions*, Pattern Recognition, vol. 29, no. 1, 1996, pp. 51-59.

[11] T. Ojala, M. Pietikäinen, and T. Maenpää, *Multiresolution Gray Scale and Rotation Invariant Texture Classification with Local Binary Patterns*, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 24, issue. 7, 2002, pp. 971-987.

[12] J.Y. Ramel, S. Busson, and M. L. Demonet, *AGORA: the Interactive Document Image Analysis Tool of the BVH Project*, Proceedings of the Second International Conference on Document Image Analysis for Libraries (DIAL), 2006, pp. 145-155.

[13] M. Seuret, *Printed Content & Handwritten Notes Discrimination in Scanned Documents*, Master Project, University of Fribourg, Department of Informatics, 2013.

[14] H. Wei, M. Baechler, F. Slimane, and R. Ingold, *Evaluation of SVM, MLP and GMM Classifiers for Layout Analysis of Historical Documents*, International Conference on Document Analysis and Recognition (ICDAR), 2013, pp. 1220-1224.

[15] L. Yu and H. Liu, *Feature selection for high-dimensional data: A fast correlation-based filter solution*, Proceedings of the Twentieth International Conference on Machine Learning, AAAI Press, 2003, pp. 856-863.