Recognition and Enrichment of Archival Documents

D6.2
Binarization and Image Enhancement Tools P2

Ioannis Pratikakis, Konstantinos Zagoris DUTH
Basilis Gatos, NCSR

Distribution:

http://read.transkribus.eu/

READ
H2020 Project 674943

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 674943

Project ref no.	H2020 674943
Project acronym	**READ**
Project full title	**Recognition and Enrichment of Archival Documents**
Instrument	H2020-EINFRA-2015-1
Thematic Priority	EINFRA-9-2015 - e-Infrastructures for virtual research environments (VRE)
Start date / duration	01 January 2016 / 42 Months

Distribution	Public
Contractual date of delivery	31.12.2017
Actual date of delivery	04.12.2017
Date of last update	
Deliverable number	D6.2
Deliverable title	Binarization and Image Enhancement Tools P2
Type	Demonstrator
Status & version	Public & Version 3
Contributing WP(s)	WP6
Responsible beneficiary	DUTH
Other contributors	
Internal reviewers	
Author(s)	Ioannis Pratikakis, Konstantinos Zagoris DUTH Basilis Gatos, NCSR
EC project officer	
Keywords	Document Image Binarization, Ground Truth, Document Image Enhancement

Table of Contents

Executive Summary

This deliverable report on the achievements concerning the tasks of document image binarization at the end of the second year of the READ project. This year DUTH concentrated on increasing the accuracy of previous binarization algorithm (M12 deliverable).

1. DUTH Framework

1.1. Binarization

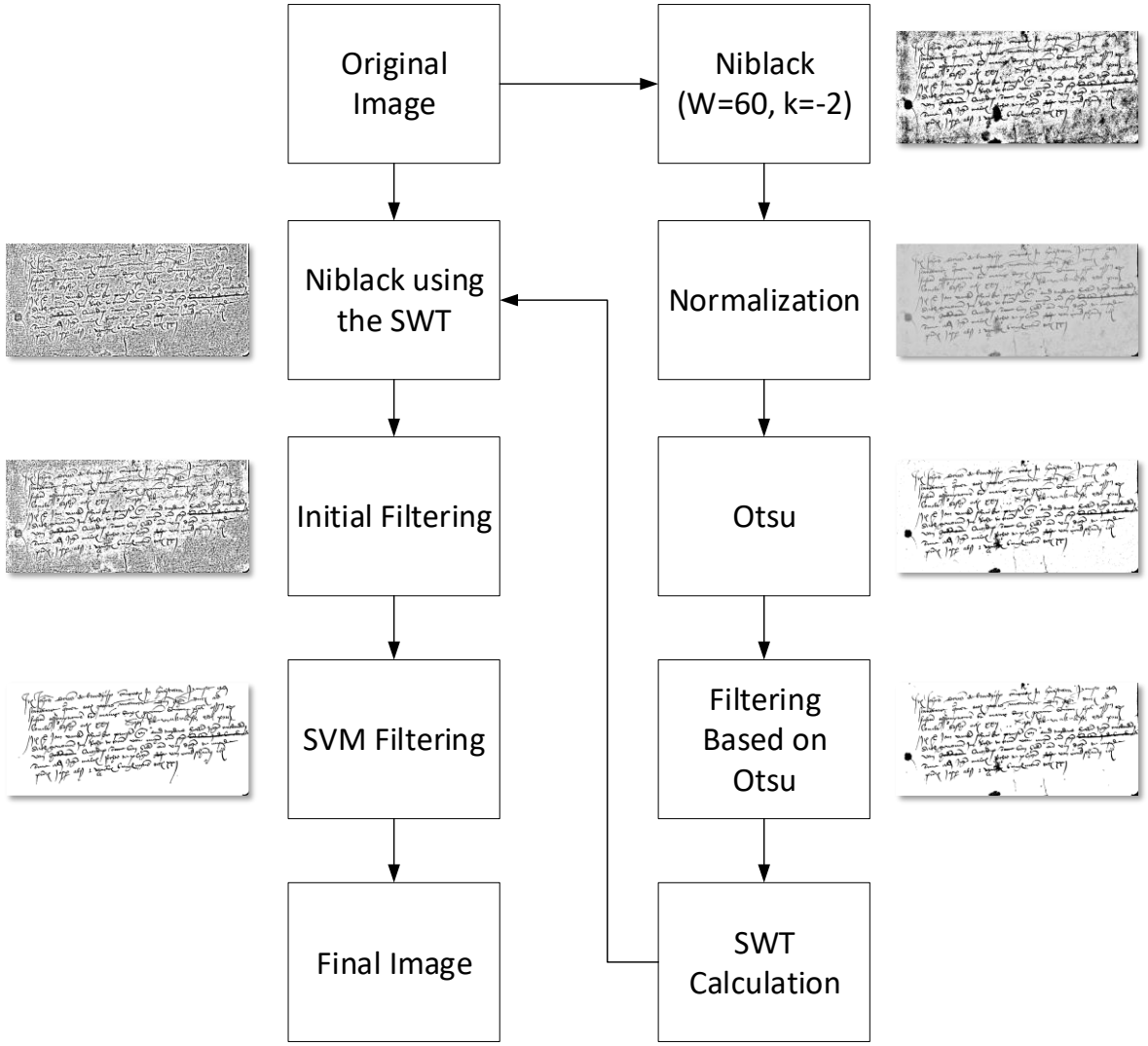

Figure 1: Architecture Diagram of the M24 Binarization method

During the second year, major modifications were applied to the M12 Binarization (D6.1) method in order to create the M24 Binarization method. Figure 1 shows its architectural structure.

Briefly, these changes are as follows:

1. New inpainting method for the creation of normalized image. [BER2001]
2. New approach to the stroke width calculation using morphological operations
3. New filtering step based on the dilated image of the original image to reduce false positive connected components
4. New machine learning-based filtering step taking into account texture information for further noise removal.
5. Removal of the combination stage Otsu-Niblack appearing in method M12

Moreover, some characteristics of the enhancement technique are incorporated into the binarization algorithm therefore it is unnecessary to use it.

Figure 2 shows some specific problems that were solved using the DUTH M24 Binarization method.

In our comparative study, we have considered a binarization method which was developed in the frame of "tranScriptorium" project that is based on the method of Ntirogiannis et al. [NTI2014] adapted to handwritten document images. This method was also used in the evaluation performed during Y1 of the project (see previous year's deliverable D6.1). This algorithm is presented in three variations, denoted as 'NCSR – method (i)-(iii)', respectively.

Table 1 shows the evaluation results using the datasets from the ICFHR2016 Handwritten Document Image Binarization Contest (H-DIBCO 2016) [PRA2016] and ICDAR2017 Competition on Document Image Binarization (DIBCO 2017) [PRA2017]. Also, the performance of the first ranked method in these competitions is also presented, denoted as '#1 Competition Method'**.**

The DUTH M12 method although performed very well on H-DIBCO2016 dataset, it did not achieve similar performance on DIBCO2017 dataset. This is because it cannot handle very well the bleed-through characters. Unfortunately, the distinction between faint characters that should be preserved and bleed-through characters that should be omitted is a very difficult and in some cases near impossible task. The DUTH M24 method tries to solve this problem by applying a filtering step in which a HOG-based SVM classifier is used to make a distinction between faint and bleed-through characters. This resulted in a balanced performance for H-DIBCO 2016 dataset between the top achievements of the DUTH and NCSR approaches and resulted in the top performance of DUTH M24 approach for DIBCO 2017 dataset compared to all other approaches produced in READ. At Table 1, it is also shown the results of the first ranked algorithm for DIBCO 2017 competition which achieves the best performance. It should be mentioned that this top ranked approach relies upon deep learning models which is the motivation for incorporating the use of these type of models in our plans for the Y3 of the project.

The DUTH M24 binarization method is developed in C++11 and is available at github under LGPL-3.0:

https://github.com/Transkribus/VCG-DUTH-ImageBinarizationTool

Table 1: Quantitative performance evaluation of M12 and M24 document image binarization methods for H-DIBCO2016 and DIBCO2017 datasets.

	H-DIBCO2016				DIBCO2017			
Framework	FM	pseudo -FM	PSNR	DRD	FM	pseudo -FM	PSNR	DRD
DUTH M12	90.62	90.95	19.04	**3.86**	83.80	84.04	15.52	8.52
DUTH M12 ***(/w enhancement)***	86.64	91.39	17.67	4.67	82.44	88.30	15.17	712
DUTH M24	88.81	**91.62**	18.27	4.47	**85.2**	**87.49**	**16.37**	**5.32**
NCSR – method (i)	88.67	87.61	18.42	5.80	79.27	77.64	14.03	13.51
NCSR – method (ii)	**90.84**	89.79	**19.31**	4.00	82.94	81.23	15.18	9.81
NCSR – method (iii)	84.40	83.49	16.52	9.29	76.7	75.17	13.15	15.73
#1 Competition Method	*87.61*	*91.28*	*18.11*	*5.21*	***91.04***	***92.86***	***18.28***	***3.40***

Figure 2: Example results that show qualitative differences between M12 and M24 algorithm

2. References

[BER2001] Bertalmio, M., Bertozzi, A. L., & Sapiro, G. (2001). Navier-stokes, fluid dynamics, and image and video inpainting. In *Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on* (Vol. 1, pp. I-I). IEEE.

[NTI2014] K. Ntirogiannis, B. Gatos and I. Pratikakis, "A Combined Approach for the Binarization of Handwritten Document Images", *Pattern Recognition Letters - Special Issue on Frontiers in Handwriting Processing*, vol. 35, no.1, pp. 3-15, Jan. 2014.

[PRA2016] Pratikakis, I., Zagoris, K., Barlas, G., & Gatos, B. (2016, October). ICFHR2016 Handwritten Document Image Binarization Contest (H-DIBCO 2016). In *Frontiers in Handwriting Recognition (ICFHR), 2016 15th International Conference on* (pp. 619-623). IEEE.

[PRA2017] Pratikakis, I., Zagoris, K., Barlas, G., & Gatos, B. (2017, November). ICDAR 2017 Document Image Binarization Contest (DIBCO 2017). In *Document Analysis and Recognition (ICDAR), 2017 14th IAPR International Conference on* (pp. 1471-1476). IEEE.