ATTENTION-BASED AUTOENCODER-LIKE ORTHOGONAL NMF FOR MULTISPECTRAL IMAGE DECOMPOSITION

*Thomas Olive** *Abderrahmane Rahiche** *Mohamed Cheriet*

Synchromedia Lab, École de Technologie Supérieure (ÉTS), Montreal, Canada

ABSTRACT

This paper introduces a novel unsupervised model for blind decomposition of multispectral (MS) images. Our approach employs a convolutional autoencoder to construct a Nonnegative Matrix Factorization (NMF) framework, allowing the model to capture complex spatial features in the data effectively. We incorporate a visual attention mechanism to enhance further the model's ability to focus on spatially significant information during the decomposition process. An orthogonality constraint is applied to improve the separation and uniqueness of the extracted components. Extensive experiments on real-world MS datasets of historical documents demonstrate the models superior performance in decomposing complex scenes. Additionally, our proposed model achieves significant model compression, reducing the number of parameters by half, as the decoder relies solely on the number of spectral bands and components. The code is available at `https://github.com/arahiche/Attention-based-AE-ONMF`.

***Index Terms*—** Autoencoder, Nonnegative Matrix Factorization, Multispectral Images, Attention, Orthogonality

1. INTRODUCTION

Blind or unsupervised decomposition of multispectral (MS) and hyperspectral (HS) images is a fundamental problem in multi-band image analysis. It aims to separate an MS/HS image into its constituent components, typically materials reflectance and abundance maps from mixed pixel data, without prior knowledge of their unique spectral signatures or their mixture coefficients. MS/HS images capture data at different wavelengths across the electromagnetic spectrum, providing rich information for various applications such as remote sensing, agriculture, food inspection, and cultural heritage analysis. However, the complexity of these images, including variations in illumination and atmospheric conditions, poses significant challenges to their effective analysis and decomposition.

Nonnegative Matrix Factorization (NMF) has emerged as a powerful technique for MS/HS image decomposition. NMF seeks to decompose a nonnegative matrix $\mathbf{X} \in \mathbb{R}+^{m\times n}$ into the product of two nonnegative matrices, an endmember matrix $\mathbf{M} \in \mathbb{R}+^{m\times r}$ and an abundance maps matrix $\mathbf{A} \in \mathbb{R}_{+}^{r\times n}$, such that $\mathbf{X} \approx \mathbf{MA}$ and $r \ll \min\{n, m\}$, where m is the number of bands, n is the total number of pixels of MS images, and k is the number of objects in the image scene. The NMF problem solves

$$\min_{\mathbf{M},\mathbf{A}\geq 0} \|\mathbf{X} - \mathbf{MA}\|_F, \tag{1}$$

where $\|.\|$ is the Frobenius norm.

Due to its unique ability to preserve the nonnegativity inherent in physical reflectance and the abundance values in MS and HS data, NMF has gained a lot of attention in this field. However, the linear model used in traditional NMF, i.e., Eq. (1), ignores the complex spectral-spatial relationships present in spectral data, which limits its ability to capture the non-linear interactions between spectral bands and spatial features. These interactions often stem from complex material properties, varying environmental conditions, and other non-linear phenomena that cannot be adequately modeled by a purely additive approach. To address this limitation, several extensions have been proposed, including kernel-based methods [1] and probabilistic reformulations [2] of the problem, which introduce non-linear transformations to better capture the underlying data structure. On the other hand, despite the inherently 3D structure of MSI and HSI data, NMF methods simplify the problem by employing a 2D representation. This is typically achieved through 3D-2D reshaping, which flattens the spatial and spectral dimensions of the data cube into a 2D matrix. While this approach simplifies computation, it fundamentally disrupts the spatial correlations between pixels, thereby losing critical spatial information that is essential for accurate material discrimination.

Inspired by the success of deep learning methods, the analogous functionality between NMF and autoencoders (AE) has motivated recent research to combine the strengths of both approaches and develop autoencoder-like NMF models [3, 4, 5]. Despite differences in their underlying mechanisms, both techniques aim to decompose high-dimensional data into compact and low-dimensional representations. By leveraging this analogy, a combined model aims to maintain the interpretability and non-negativity properties of NMF while overcoming its limitations through the non-linear capabilities of AE, leading to more robust models that can potentially capture more complex patterns in the data. Sev-

*Authors contributed equally. The authors thank NSERC Canada for its financial support.

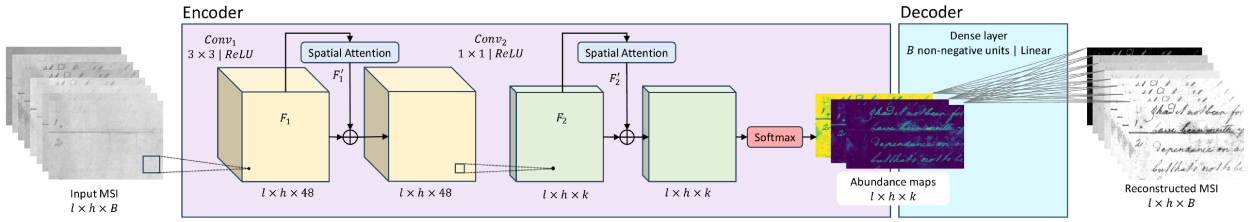


Fig. 1. Pipeline of the proposed model.

eral approaches have been proposed recently to address the problem of HS unmixing, such as DNAE [6] and [5]. However, these AE-NMF models still rely on 2D matrix form, take pixel spectra as input, and use fully connected layers to form the structure of the encoder, which processes each pixel independently, ignoring the spatial information and neighbor relationships among pixels. Later, this limitation has been addressed in [7], by a convolutional structure that relies on 3D cube representation. Convolutional AE-NMF models [8] are more able to learn complex spatial features by applying filters that capture local patterns and textures. Nevertheless, most existing approaches focus on the HS unmixing problem. To the best of our knowledge, this is the first work that addresses the problem of MS document image decomposition using an AE-NMF model.

1.1. Contribution

The contributions of this study are as follows: 1) we address the decomposition of multispectral document images, which is less covered in the literature. 2) we employ a convolutional structure to form the encoder, which handles nonlinearity in MS data and enhances feature extraction capabilities, and a fully connected (FC) decoder to regenerate MS data, which reduces its number of parameters from several thousand to $m \times k$ only, which is a very interesting compression. 3) An attention mechanism is incorporated to allow the model to focus on relevant spatial features during the decomposition process. 4) Experiments conducted on real-world MS document images show that the model performs well, particularly in scenarios involving severely degraded images.

2. THE PROPOSED VNAE-ONMF MODEL

The proposed visual attention network based model, called VNAE-ONMF, consists of an encoder-decoder structure with attention mechanisms and an orthogonality constraint to effectively process high-dimensional spectral data. The encoder extracts abundance maps from the input cube of MS images, representing the relative proportions of spectral endmembers in each pixel, while the decoder reconstructs the output image by combining these abundance maps with learned endmember spectra. An overview of the full architecture is illustrated in Fig. 1.

2.1. Encoder

The encoder consists of two cascaded 2D convolutional layers, each followed by a ReLU (Rectified Linear Unit) activation function. The first convolutional layer, Conv1, employs 48 filters of size 3×3, enabling the capture of local spatial relationships. The second layer, Conv2, uses k feature maps (number of components) with a 1×1 filter size, typically serving for dimensionality reduction and feature aggregation, a process that combines information from multiple channels into a more compact representation. The two ReLU activations introduce non-linearity, enhancing the model's capacity to learn complex data representations.

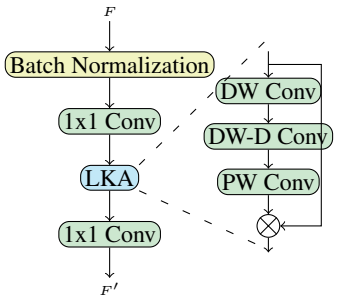


Fig. 2. Structure of the Spatial Attention block (SAB). $\otimes$ denotes an element-wise product operator.

The Spatial Attention block in Fig. 2 is composed of a Batch Normalization (BN) layer to stabilize learning, followed by an initial 1×1 convolution layer. After this convolution, we incorporated the Visual attention network [9], also called Large-Kernel-Attention (LKA), which serves as the backbone of the block. The LKA consists of three successive convolution layers: a 5×5 Depth-Wise (DW) convolution for local information, a 1×1 Dilated Depth-Wise (DW-D) convolution with a dilation rate of 3 for larger information, and a 1×1 Point-Wise (PW) convolution for cross-channel information mixing. This structure allows the LKA to capture both local and long-range spatial dependencies efficiently. The block concludes with an additional 1×1 convolution layer, used for projecting attention-weighted features back to the original feature space. As shown in Fig. 2, the SAB module can be formulated as:

$$F_N = \text{Conv}_{1\times1}(BN(F)) \tag{2}$$

$$\text{Attention} = \text{Conv}_{PW}(\text{Conv}_{DW-D}(\text{Conv}_{DW}(F_N))) \tag{3}$$

$$\text{Output} = \text{Conv}_{1\times1}(\text{Conv}_{PW} \otimes F_N) \tag{4}$$

The encoder concludes with a softmax function applied after the 1×1 convolution layer, ensuring compliance with the Abundance Non-negativity Constraint (ANC), i.e., $\mathbf{A} \geq 0$.

2.2. Decoder

The decoder consists of a single FC layer, which serves as an analog to the endmember matrix in conventional NMF. The weights of this layer effectively represent the spectral signatures of the endmembers, mapping the latent space representation back to the original dimensions of the images. A non-negativity constraint is applied to the decoder weights to reconstruct the original data representation.

2.3. Loss function

The Mean Squared Error (MSE) is used for the loss function of our model to optimize the quality of the reconstruction. MSE quantifies the average squared difference between corresponding elements of the original input image and the reconstructed output, encouraging the model to produce reconstructions that closely match the input data. The MSE between elements of two matrices, x and $\hat{x}$, is written as :

$$\mathcal{L}_{MSE} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{x}_i)^2, \tag{5}$$

where x is the input spectra of the i^{th} pixel and x_i is the corresponding output.

As for the orthogonality constraint, enforcing orthogonality on the abundance matrix $\mathbf{A}$ has been shown to be particularly relevant in the literature [10]. Therefore, we adopt a similar approach, proposing a more flexible version of this constraint. To make it compatible with the Abundance Sum-to-One Constraint (ASC) condition, i.e., $\sum_{i=1}^{n} a_{ij} = 1$, only off-diagonal values of the $\mathbf{A}\mathbf{A}^T$ matrix are considered in the loss function formula (6).

$$\mathcal{L}_{orth} = \|\mathbf{A}\mathbf{A}^T - \text{Diag}(\text{diag}(\mathbf{A}\mathbf{A}^T))\|_2, \tag{6}$$

where diag is the diagonal operator that extracts the diagonal elements from matrix $\mathbf{A}\mathbf{A}^T$, and Diag is the operator that forms a diagonal matrix. This reduces linear correlation between extracted images, so that a single element does not appear on several of them. The total loss for patch β^i writes

$$\mathcal{L}_{\beta} = \frac{1}{|\beta|}\sum_{x\in\beta}\mathcal{L}_{MSE}(x,\hat{x}) + \lambda\mathcal{L}_{orth}(\mathbf{A}), \tag{7}$$

where λ is a penalty prameter.

3. EXPERIMENTS

To evaluate our approach, we focus on the decomposition of MS images of historical documents, which are of particular interest to us. Moreover, this type of data has received limited exploration in the literature.

3.1. Datasets

For our evaluation, we used three different datasets: MSTEx1 [11], MSTEx2 [11], and MSBin [12]. All these datasets were specifically designed for text extraction and binarization from multispectral document images. Although other objects, such as stamps, may be present in the scenes, the ground truth (GT) is provided for text only.

The two MSTEx datasets consist of 30 images of historical manuscripts from the 17th to 20th centuries. Each image is captured across 8 spectral bands from 340nm to 1100nm, using a Chroma X3 KAF 6303E (Kodak) sensor, with a resolution of 6 megapixels ($9 \times 9\mu$ m). The MSBin dataset[1] contains MS images captured with a Phase One IQ260 achromatic camera at a resolution of 60 megapixels, across 12 narrow-band spectral ranges from 365nm to 940nm. The images are taken from two manuscripts: Bitola-Triodion ABAN 38 (Book BT) and Enina-Apostolus NBMK 1144 (Book EA), the latter being severely degraded. Only the test set is used in our experiments.

3.2. Baseline Methods

Since the abundance maps generated by NMF are not binary, a binarization step is applied so they can be compared with the provided binary GT images. Howe's method [13] is adopted for this post-processing step as it is one of the best-known techniques for text image binarization. In total, we compare 8 different models, namely: MA-ONMF [10], a bi-orthogonal NMF model, GMM [14], a Gaussian mixture-based model, FCN, a fully connected network [15], CNN [12], the baseline model used for MSBin, SKKHM [16], a Spatial kernel K-harmonic means clustering method, and the versions of the Adaptive Coherence Estimator method (ACE) based on matched filters, ACE1 [17] and ACE2 [18].

3.3. Metrics

We assessed the performance of the model on text-extraction tasks using five common quantitative metrics, namely F-measure (FM) (also known as F1-score), Distance Reciprocal Distortion (DRD), Negative Rate Metric (NRM), Peak Signal to Noise Ratio Accuracy (PSNR), and pseudo-F-measure (p-FM)[2]. The FM, p-FM, and PSNR are expressed as a percentage, while DRD and NRM values range from 0 to 1. A higher value (↑) of FM, p-FM, and PSNR indicates a better extraction quality, whereas lower (↓) NRM and DRD values suggest superior performance.

3.4. Model Parameters

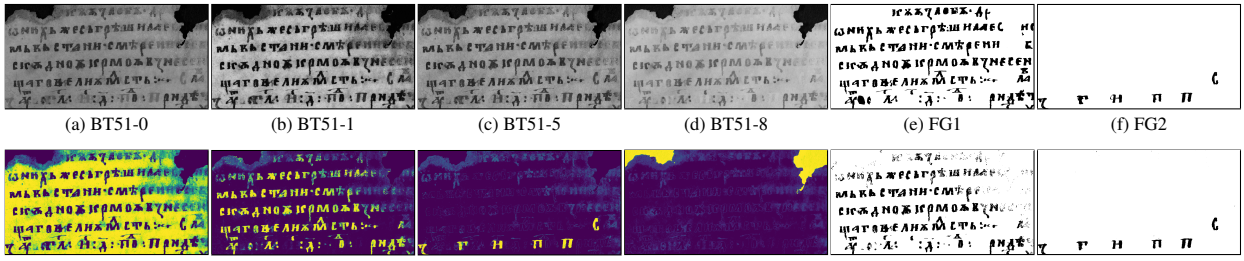
Our model is trained on N patches (250 patches of size 40×40 each), randomly chosen from an input MS image. The initial abundance maps ($\mathbf{A}$) and endmember matrices ($\mathbf{M}$) are obtained at the end of the training phase and are used to generate the final version from the full image. We used the Root Mean Square Propagation (RMSprop) as an adaptive learning rate optimization algorithm and we set the initial learning rate value to 10^{-3}, the batch size is set to 15, and we found that $\lambda = 10^{-5}$ gives better results.

3.5. Results

In this section, we present our qualitative and quantitative results, including abundance maps of the decomposed outputs

[1] https://doi.org/10.5281/zenodo.3257365

[2] For the sake of space, we omit the definitions of these metrics. Readers can refer to [19] for further details.



(a) BT51-0 (b) BT51-1 (c) BT51-5 (d) BT51-8 (e) FG1 (f) FG2

(g) Paper (h) Text 1 (i) Text 2 (j) Degradation (k) Binary text 1 (l) Binary text 2

Fig. 3. Abundance maps and the corresponding bianry images extracted from image BT51 using our approach. (a-d) are the MS images, (e) and (f) are the provided foregrounds, (g)-(j) the abundance maps of the extracted components, and (k)-(l) are the binarized text components.

and numerical results from the binarization process (text extraction). The first experiment is conducted on the two MSTEx datasets. The numerical results are reported in Table1 and Table 2.

Table 1. Average binarization results of different methods on the MSTEx-2 dataset.

Metric	Howe	SKKHM	GMM	MA-ONMF	Proposed
FM(%) ↑	70.41	62.68	82.05	83.86	**84.29**
DRD($\times 10^{-2}$) ↓	8.58	17.16	4.54	3.72	**3.58**
NRM($\times 10^{-2}$) ↓	12.06	16.33	8.67	8.92	**7.67**
PSNR↑	15.03	13.29	16.97	-	**21.33**

As shown in Table 2, the proposed model outperforms other approaches across all metrics, demonstrating its effectiveness in text extraction.

Table 2. Average binarization results of different methods on the MSTEx-1 dataset.

Metric	Howe	SKKHM	GMM	MA-ONMF	Proposed
FM(%) ↑	81.83	79.86	79.82	**86.32**	85.42
DRD($\times 10^{-2}$) ↓	6.34	5.48	5.52	**3.46**	3.75
NRM($\times 10^{-2}$) ↓	5.23	10.34	11.61	**6.11**	6.58
PSNR↑	15.07	15.37	15.42	-	**17.39**

This experiment demonstrates the superiority of two methods, MA-ONMF and Proposed, over other baseline approaches. The MA-ONMF method exhibits the best overall performance, with the highest F-Measure of 86.32% and lowest DRD of 3.46, attributed to its robust orthogonality technique via direct optimization on the Stiefel manifold. The Proposed method shows comparable performance, particularly standing out with the highest PSNR value of 17.39.

Results on the MSBin dataset, show that FCN and CNN demonstrate competitive performance, particularly in the FM metric. Our model outperforms all baselines in PSNR for both EA and BT images, although its performance declines in other metrics. Notably, our model performs exceptionally well on BT images. A deeper analysis revealed difficulties with only four images: EA56, EA61, EA67, and EA68, which are severely degraded. When these images are excluded, the model achieves high scores (in blue) in terms of FM: 80.86, p-FM: 88.58, PSNR: 16.43, DRD: 18.46, and NRM: 11.78, making it highly competitive with the baseline results.

Table 3. Test extraction and differentiation on MSBin dataset.

Method	FM(%) ↑	p-FM(%) ↑	PSNR↑	DRD↓	NRM↓
Ours (EA only)	53.18	69.60	12.44	64.42	28.77
Ours (BT only)	**90.35**	**94.83**	**18.47**	10.28	5.95
Ours (BT + EA)	74.55	84.11	**15.90**	33.29	15.65
CNN [12]	88.20	-	-	-	5.71
ACE1 [17]	81.28	81.18	13.28	22.03	-
ACE 2 [18]	81.25	81.36	13.27	20.80	-
GMM [14]	80.00	80.28	13.18	20.35	-
FCN [15]	**89.39**	**90.89**	15.17	**9.91**	-

For a qualitative assessment, Fig. 3 showcases one of the obtained results. The MS image BT51 features two main texts written in different inks, along with a degradation where the paper does not cover the entire scene. Our approach successfully decomposed the image, effectively separating the paper, degradation, and the two distinct texts from each other.

Finally, it is important to highlight that the proposed approach achieves a significant reduction in the total number of parameters, as the decoder requires only $m \times k$ parameters, effectively reducing the number of parameters by half compared to the traditional full decoder block. This substantial reduction translates into lower computational requirements.

Table 4. Number of parameters and model size of our model, $m = 8$ and $k = 4$.

Component	**Number of Parameters**	**Model Size (KB)**
Encoder	12324	48.14
Decoder	32	0.13
Total	12356	48.27

4. CONCLUSION

The proposed model offers significant advancements in the analysis and interpretation of complex multispectral data by effectively handling nonlinearity and enhancing feature extraction through an autoencoder architecture. The integration of attention mechanisms enables focused analysis of key spatial features, while convolutional layers preserve spatial information and account for inherent nonlinearity. Additionally, parameter compression in the decoder reduces model complexity, making the approach more efficient without compromising performance. The models strong results on real-world MS document images, particularly in decomposing severely degraded scenes into distinct constituent elements, underscore its practical applicability and efficiency.

5. REFERENCES

[1] Abderrahmane Rahiche and Mohamed Cheriet, "Kernel orthogonal nonnegative matrix factorization: Application to multispectral document image decomposition," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 3275–3279.

[2] Abderrahmane Rahiche and Mohamed Cheriet, "Variational bayesian orthogonal nonnegative matrix factorization over the stiefel manifold," *IEEE Transactions on Image Processing*, vol. 31, pp. 5543–5558, 2022.

[3] Ying Qu, Rui Guo, and Hairong Qi, "Spectral unmixing through part-based non-negative constraint denoising autoencoder," *2017 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, pp. 209–212, 2017.

[4] Yuanchao Su, Andrea Marinoni, Jun Li, Antonio J. Plaza, and Paolo Gamba, "Nonnegative sparse autoencoder for robust endmember extraction from remotely sensed hyperspectral images," *2017 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, pp. 205–208, 2017.

[5] Frosti Palsson, Jakob Sigurdsson, Johannes R. Sveinsson, and Magnus Orn Ulfarsson, "Neural network hyperspectral unmixing with spectral information divergence objective," *2017 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, pp. 755–758, 2017.

[6] Yuanchao Su, Jun Li, Antonio Plaza, Andrea Marinoni, Paolo Gamba, and Somdatta Chakravortty, "Daen: Deep autoencoder networks for hyperspectral unmixing," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 7, pp. 4309–4321, 2019.

[7] Burkni Palsson, Magnus Orn Ulfarsson, and Johannes R. Sveinsson, "Convolutional autoencoder for spatial-spectral hyperspectral unmixing," *IGARSS 2019 - 2019 IEEE International Geoscience and Remote Sensing Symposium*, pp. 357–360, 2019.

[8] Xia Xu, Xinyu Song, Tao Li, Zhenwei Shi, and Bin Pan, "Deep autoencoder for hyperspectral unmixing via global-local smoothing," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–16, 2022.

[9] Meng-Hao Guo, Cheng-Ze Lu, Zheng-Ning Liu, Ming-Ming Cheng, and Shi-Min Hu, "Visual attention network," *Computational Visual Media*, vol. 9, no. 4, pp. 733–752, 2023.

[10] Abderrahmane Rahiche and Mohamed Cheriet, "Blind decomposition of multispectral document images using orthogonal nonnegative matrix factorization," *IEEE Transactions on Image Processing*, vol. 30, pp. 5997–6012, 2021.

[11] Rachid Hedjam, Hossein Ziaei Nafchi, Reza Farrahi Moghaddam, Margaret Kalacska, and Mohamed Cheriet, "Icdar 2015 contest on multispectral text extraction (ms-tex 2015)," in *2015 13th International Conference on Document Analysis and Recognition (ICDAR)*. IEEE, 2015, pp. 1181–1185.

[12] Fabian Hollaus, Simon Brenner, and Robert Sablatnig, "Cnn based binarization of multispectral document images," in *2019 International Conference on Document Analysis and Recognition (ICDAR)*. IEEE, 2019, pp. 533–538.

[13] Nicholas R Howe, "Document binarization with automatic parameter tuning," *International journal on document analysis and recognition (ijdar)*, vol. 16, pp. 247–258, 2013.

[14] Fabian Hollaus, Markus Diem, and Robert Sablatnig, "Multispectral image binarization using gmms," in *2018 16th International Conference on Frontiers in Handwriting Recognition (ICFHR)*. IEEE, 2018, pp. 570–575.

[15] Jonathan Long, Evan Shelhamer, and Trevor Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3431–3440.

[16] Qi Li, Nikolaos Mitianoudis, and Tania Stathaki, "Spatial kernel k-harmonic means clustering for multispectral image segmentation," *IET Image Processing*, vol. 1, no. 2, pp. 156–167, 2007.

[17] Fabian Hollaus, Markus Diem, and Robert Sablatnig, "Binarization of multispectral document images," in *Computer Analysis of Images and Patterns: 16th International Conference, CAIP 2015, Valletta, Malta, September 2-4, 2015, Proceedings, Part II 16*. Springer, 2015, pp. 109–120.

[18] Markus Diem, Fabian Hollaus, and Robert Sablatnig, "Msio: Multispectral document image binarization," in *2016 12th IAPR Workshop on Document Analysis Systems (DAS)*. IEEE, 2016, pp. 84–89.

[19] Chris Tensmeyer and Tony Martinez, "Historical document image binarization: A review," *SN Computer Science*, vol. 1, no. 3, pp. 173, 2020.