Document Image Binarization Algorithm Combining Background Estimation and U-Net (Postprint)

Authors: Xiong Wei, Xinrui Wang, Wang Juan, Liu Min, Zeng Chunyan

Date: 2019-01-03T00:00:00+00:00

Abstract

To address various degradation factors present in low-quality document images, such as page stains, ink bleed-through, and background texture, this paper proposes a document image binarization algorithm that integrates background estimation with a U-shaped convolutional neural network (U-Net). The algorithm first performs image contrast enhancement, then estimates the document image background through morphological closing operations, and employs a fully convolutional network—specifically, U-Net—to conduct foreground-background segmentation on the background-subtracted image. Finally, a global optimal thresholding method is applied to obtain the final binary image. Experimental results show that in the 2016 and 2017 International Document Image Binarization Competitions, the proposed algorithm achieves performance improvements of up to 5.58%, 2.47%, 0.86 dB, and 1.19% in terms of F-measure (FM), pseudo F-measure (p-FM), peak signal-to-noise ratio (PSNR), and distance reciprocal distortion (DRD), respectively, compared to the second-best performing classical algorithm.

Full Text

Preamble

Document Image Binarization Algorithm Based on Background Estimation and U-Net

Xiong Weia,b, Wang Xinruia, Wang Juana,b, Liu Mina,b, Zeng Chunyana,b

aSchool of Electrical & Electronic Engineering; bHubei Collaborative Innovation Center for High-efficiency Utilization of Solar Energy, Hubei University of Technology, Wuhan 430068, China

Abstract: Degraded document images suffer from various degradation factors such as page stains, ink bleed-through, and background texture. We propose a novel document image binarization algorithm that integrates background estimation with a U-shaped convolutional neural network (U-Net). The algorithm first enhances image contrast, then estimates the document background through morphological closing operations. A fully convolutional network, namely U-Net, is employed to segment foreground from background in the background-subtracted image. Finally, a global optimal thresholding method is applied to obtain the binary result. Experimental results on the 2016 and 2017 International Document Image Binarization Contests demonstrate that our algorithm achieves performance improvements of up to 5.58% in F-measure (FM), 2.47% in pseudo F-measure (p-FM), 0.86 dB in peak signal-to-noise ratio (PSNR), and 1.19% in distance reciprocal distortion (DRD) compared to the second-best classical algorithms.

Keywords: document image binarization; contrast enhancement; morphological closing operation; U-Net; global optimal thresholding

0 Introduction

Image binarization serves as a crucial preprocessing step for document image recognition and analysis, with applications in ancient document restoration and signature verification [1]. Due to physical conditions and human factors, low-quality document images exhibit complex background characteristics such as page stains [2], resulting in minimal differentiation between textual information and background. Consequently, binarization of low-quality document images remains highly challenging [3].

Document image binarization algorithms are categorized into global thresholding, local thresholding, and hybrid methods [4]. Global thresholding applies a fixed threshold to all pixels, offering low computational complexity but often causing text loss in images with complex backgrounds. The classic Otsu algorithm [5] exemplifies this approach. Local thresholding determines thresholds for each pixel through convolution operations with sliding windows. Wolf [6] performs local binarization using normalized contrast and the standard deviation and mean gray value within neighborhoods. Other local methods include Sauvola [7] and Niblack [8] algorithms, which are more adaptable but whose performance heavily depends on window size.

Su et al. [9] normalized local gray values to suppress uneven background illumination (LMM algorithm), though this creates hollow characters at edge regions. Lu et al. [10] detected character edges based on pixel differences within neighborhoods and estimated stroke width for binarization (BESE algorithm), but showed poor ability to suppress background stains. Howe achieved image segmentation through graph-cut energy function minimization [11] and optimized structural parameters [12], yet such methods lose stroke details in low-contrast images. Mesquita [13] proposed a human visual model to distinguish text from

background pixels using energy function minimization and I/F-Race methods. Kligler [14] removed estimated backgrounds based on brightness variations and applied graph-cut algorithms, but misclassified bleed-through ink as characters. Tensmeyer [15] combined relative darkness features with energy functions using a five-layer fully convolutional neural network, which suppressed background stains but caused stroke discontinuities. Additional hybrid methods include parameter optimization [16], classifier-based [17], and clustering approaches [18]. This paper integrates background estimation with a U-shaped convolutional neural network for document image binarization, with experimental validation demonstrating its superior performance.

1 Document Image Binarization Algorithm Integrating Background Estimation and U-Net

1.1 Algorithm Flow

The proposed method is illustrated in [Figure 1: see original paper]. The algorithmic flow comprises four main stages: (a) grayscale conversion of color document images using weighted averaging to obtain grayscale image f_{gray}, followed by contrast enhancement; (b) background estimation of the contrast-enhanced image f_{eq} via morphological closing operations, where structuring element size correlates with text stroke width; (c) computation of the difference image between f_{eq} and background estimate f_{bg} to produce the background-subtracted image f_{negate}; and (d) segmentation of f_{negate} through the U-Net to obtain f_{seg}, with final binary result f_{final} generated using Otsu' s method.

1.2 Background Estimation

The original document image undergoes grayscale conversion using the weighted average method shown in equation (1) to produce f_{gray} [FIGURE:2(b)].

$$f_{gray}(x,y) = 0.299 \times R(x,y) + 0.587 \times G(x,y) + 0.114 \times B(x,y)$$

where $R(x,y)$, $G(x,y)$, and $B(x,y)$ represent the red, green, and blue channel components of the image.

Linear grayscale transformation is applied to f_{gray} as shown in equation (2) to obtain f_{eq} [FIGURE:2(c)], enhancing contrast between characters and background.

$$f_{eq}(x,y) = \begin{cases} l_1 & \text{if } f_{gray}(x,y) < l_1 \\ \frac{h_2-h_1}{l_2-l_1} \times (f_{gray}(x,y) - l_1) + h_1 & \text{if } l_1 \le f_{gray}(x,y) \le l_2 \\ h_2 & \text{if } f_{gray}(x,y) > l_2 \end{cases}$$

where pixels with gray values below l_1 and above l_2 constitute 1% and 10% of the total image, respectively, and $h_1 = 0$, $h_2 = 255$.

The stroke width transform algorithm [19] computes the gradient at pixel p using the Canny operator along ray directions r to find matching points q (with gradient d_q). If d_p and d_q have opposite directions, the Euclidean distance between p and q is calculated and points between p and q are assigned value $||pq||$ (excluding pixels already assigned smaller values). If no matching point is found, ray r is discarded. The stroke width estimate SWE for image f_{eq} is computed as:

$$SWE = \frac{1}{num} \sum_{(x,y)} s(x,y), \quad s(x,y) \neq 0$$

where num counts non-zero elements in matrix $s(x,y)$.

Morphological closing operations with a circular structuring element are performed on f_{eq} to obtain the estimated background f_{bg} [FIGURE:2(d)]. The structuring element diameter d relates to stroke width as:

$$f_{bg} = (f_{eq} \oplus b) \ominus b, \quad d = SWE + \Delta d$$

where $\oplus$ denotes dilation, $\ominus$ denotes erosion, b represents the structuring element with diameter d, and Δd is an increment.

The absolute difference between f_{eq} and f_{bg} yields difference image f_{diff} [FIGURE:2(e)], which is inverted to obtain the background-subtracted image f_{negate} [FIGURE:2(f)]. Compared to the grayscale image, this suppresses background stains and facilitates character-background separation.

To determine Δd, we selected 76 images from DIBCO 2009-2014 as training data. Otsu' s algorithm was applied to background-subtracted images, with results evaluated using F-measure:

$$F = \frac{2 \times R \times P}{R + P}, \quad R = \frac{TP}{TP + FN}, \quad P = \frac{TP}{TP + FP}$$

where TP, FP, and FN represent true positive, false positive, and false negative pixel counts, respectively. Training results are shown in .

indicates that F-measure is highest when $\Delta d = 8$, demonstrating optimal background suppression, thus this value is selected.

1.3.1 U-Net Architecture

The U-shaped convolutional neural network (U-Net) [20] enables end-to-end training with minimal images and achieved excellent performance in the ISBI neuron structure segmentation challenge. We adopt U-Net for foreground-background segmentation of background-subtracted images, with architecture shown in [Figure 3: see original paper].

The U-Net consists of a contracting path and a symmetric expanding path. The contracting path comprises basic units of two 3×3 convolutions with stride 1, followed by 2×2 max pooling with stride 2, using ReLU activation functions to produce low-resolution, high-dimensional feature maps through downsampling. The expanding path upsamples these feature maps using 2×2 transposed convolutions that halve feature channels, concatenates them with corresponding layers from the contracting path, and applies two 3×3 convolutions as in the contracting path. The final layer uses a 1×1 convolution with Sigmoid activation to map input feature vectors to output probabilities:

$$\hat{y} = S(y) = \frac{1}{1 + e^{-y}}, \quad \hat{y} \in (0, 1)$$

where y is the input feature and $\hat{y}$ represents the probability of the pixel being classified as text.

The network employs logarithmic loss to reflect differences between predictions and ground truth:

$$J = -\frac{1}{m} \sum_{i=1}^{m} [y_i \ln(\hat{y}_i) + (1 - y_i) \ln(1 - \hat{y}_i)]$$

where y_i is the ground truth, $\hat{y}_i$ is the prediction, and m is the number of samples. Model parameters are updated via backpropagation.

1.3.2 Network Training

Training data consists of document images after background estimation and their corresponding ground truth images from DIBCO 2009-2014. Considering non-uniform image sizes (minimum height of 263 pixels), we cropped images into 256×256 sub-images using a sliding window with stride 214 [Figure 4: see original paper]. The model was trained on 2,027 cropped image pairs, with output images stitched according to sliding window positions.

To maximize GPU memory utilization, we set the learning rate to 10^{-4}, batch size to 1 image, and training iterations to 10. The error rate curve is shown in [Figure 5: see original paper], stabilizing between 0.013-0.014, indicating model convergence.

1.4 Global Optimal Thresholding

The U-Net output [FIGURE:6(a)] exhibits a prominent bimodal gray histogram [FIGURE:6(b)], making it suitable for global optimal thresholding. We apply Otsu' s method, which determines L gray-level components from the histogram distribution. For each threshold $k \in [0, L-1]$, we compute foreground and background pixel proportions $P_1(k)$ and $P_2(k)$:

$$P_1(k) = \sum_{i=0}^{k} p_i, \quad P_2(k) = \sum_{i=k+1}^{L-1} p_i$$

Global mean μ and class means $\mu_1(k)$, $\mu_2(k)$ are calculated as:

$$\mu = \sum_{i=0}^{L-1} ip_i, \quad \mu_1(k) = \frac{1}{P_1(k)} \sum_{i=0}^{k} ip_i, \quad \mu_2(k) = \frac{1}{P_2(k)} \sum_{i=k+1}^{L-1} ip_i$$

Between-class variance $\sigma^2(k)$ serves as separability measure:

$$\sigma^2(k) = P_1(k)[\mu_1(k) - \mu]^2 + P_2(k)[\mu_2(k) - \mu]^2 = \frac{[\mu P_1(k) - \mu(k)]^2}{P_1(k)[1 - P_1(k)]}$$

The optimal threshold k^* maximizes $\sigma^2(k)$. The final binary image [FIGURE:7(a)] preserves character completeness while suppressing background stains visible in the bottom-right of [FIGURE:6(a)], showing strong visual similarity to ground truth [FIGURE:7(b)].

2 Experimental Results

The test dataset comprises 30 images from DIBCO 2016-2017. Evaluation metrics include F-measure (FM), pseudo F-measure (p-FM), peak signal-to-noise ratio (PSNR), and distance reciprocal distortion (DRD), where higher values for the first three indicate better accuracy and lower DRD indicates lower pixel misclassification rates [21-23].

compares our algorithm with the top three algorithms from DIBCO 2016 and 2017 (TOP1, TOP2, TOP3). Our method outperforms all three across all four metrics, demonstrating superior accuracy and robustness.

presents detailed results on DIBCO 2016 data across various algorithms. Time complexity (Time) represents average processing speed per image in seconds. Except for Score, all parameters are averaged. The Score comprehensively evaluates performance across metrics as:

$$Score(i) = \frac{1}{M} \sum_{j=1}^{M} \frac{R_{ij} - \min(R(1:N,j))}{\max(R(1:N,j)) - \min(R(1:N,j))}$$

where R contains all evaluation metrics $\{FM, p\text{-}FM, PSNR, DRD\}$, N is the number of algorithms, and M is the number of metrics. Algorithms are ranked by Score.

Experiments were conducted on an NVIDIA GTX1080 8G GPU. While Time was not included in Score calculation and thus not a performance criterion,

our algorithm shows the highest FM and PSNR values, lowest DRD and Score, though with slower processing due to higher complexity.

Tensmeyer achieves high p-FM, indicating good pixel classification accuracy. Otsu' s fixed global threshold enables fast processing. Our method delivers the best FM, PSNR, and minimal DRD and Score.

shows DIBCO 2017 results, where Howe_base achieves high PSNR, indicating high similarity to ground truth. Our method attains the highest FM and p-FM, lowest DRD and Score, outperforming both traditional binarization algorithms like Howe_alg3 and Tensmeyer' s neural network model.

[Figure 8: see original paper] presents three representative test images: gutter shadow, thin strokes, and ink bleed-through, with binarization results from various algorithms. Otsu misclassifies gutter shadows and bleed-through ink as foreground. Wolf severely breaks strokes in thin and gutter shadow images. Niblack preserves text boundaries but introduces significant noise. Sauvola loses fine character details in weak strokes. LMM creates hollow strokes at image edges. BESE misclassifies bleed-through ink as characters. Howe' s algorithms [11,12] still misclassify substantial background as foreground in bleed-through documents. Kligler exhibits stroke discontinuities and poor bleed-through handling. Tensmeyer suppresses background but neglects thin strokes.

Our proposed method effectively addresses complex background interference, accurately separating textual information and demonstrating superior visual performance.

3 Conclusion

We propose a document image binarization algorithm integrating background estimation with a U-shaped convolutional neural network. Background estimation suppresses stains, U-Net classifies document pixels, and Otsu' s method refines the binary segmentation. Experiments on DIBCO 2016 and 2017 demonstrate improvements of up to 5.58% in FM, 2.47% in p-FM, 0.86 dB in PSNR, and 1.19% in DRD over the second-best classical algorithms, confirming our algorithm' s superior performance.

References

[3] Milyaev S, Barinova O, Novikova T, et al. Fast and accurate scene text understanding with image binarization and off-the-shelf OCR [J]. International Journal on Document Analysis and Recognition, 2015, 18(2): 169-182.

[4] Eskenazi S, Gomez-Kreimer P, Ogier J M. A comprehensive survey of mostly textual document segmentation algorithms since 2008 [J]. Pattern Recognition, 2017, 64(1): 1-14.

[5] Otsu N. A threshold selection method from gray-level histograms [J]. IEEE Trans on Systems, Man, and Cybernetics, 1979, 9(1): 62-66.

[6] Wolf C, Jolion J M, Chassaing F. Text localization, enhancement and binarization in multimedia documents [C]//Proc of the 16th International Conference on Pattern Recognition, Quebec: IEEE Press, 2002: 1037-1040.

[7] Niblack W. An introduction to digital image processing [M]. Englewood Cliffs: Prentice-Hall, 1986: 115-126.

[8] Sauvola J, Pietikeiinen M. Adaptive document image binarization [J]. Pattern Recognition, 2000, 33(2): 225-236.

[9] Su Bolan, Lu Shijian, Tan Chewlim. Binarization of historical document images using the local maximum and minimum [C]//Proc of the 9th IAPR International Workshop on Document Analysis Systems.NewYork: ACM Press, 2010: 159-166.

[10] Lu Shijian, Su Bolan, Tan Chewlim. Document image binarization using background estimation and stroke edges [J]. International Journal on Document Analysis and Recognition, 2010, 13(4): 303-314.

[11] Howe N R. A laplacian energy for document binarization [C]//Proc of International Conference on Document Analysis and Recognition, Beijing: IEEE Press, 2011: 6-10.

[12] Howe N R. Document binarization with automatic parameter tuning [J]. International Journal on Document Analysis and Recognition, 2013, 16(3): 247-258.

[13] Mesquita R G, Silva R M, Mello C A, et al. Parameter tuning for document image binarization using a racing algorithm [J]. Expert Systems with Applications, 2015, 42(5): 2593-2603.

[14] Kligler N, Katz S, Tal A. Document enhancement using visibility detection [C]// Proc of the Conference on Computer Vision and Pattern Recognition, Salt: IEEE Press, 2018: 2374-2382.

[15] Tensmeyer C, Martinez T. Document image binarization with fully convolutional neural networks [J]. arXiv preprint arXiv: 170803276, 2017.

[16] Westphal F, Grahn H, Lavesson N. Efficient document image binarization using heterogeneous computing and parameter tuning [J]. International Journal on Document Analysis and Recognition, 2018, 21 (1): 41-58.

[17] Ahmadi E, Azimifar Z, Shams M, et al. Document image binarization using a discriminative structural classifier [J]. Pattern Recognition Letters, 2015, 63(1): 36-42.

[18] Jana P, Ghosh S, Bera S K, et al. Handwritten document image binarization: an adaptive K-means based approach [C]// Proc of Conference on Calcutta Conference Lalit: IEEE Press, 2017: 226-230.

[19] Yin Xucheng, Pei Weiyi, Zhang Jun, et al. Multi-orientation scene text detection with adaptive clustering [J]. IEEE Trans on Pattern Analysis and

Machine Intelligence, 2015, 37(9): 1930-1937.

[20] Ronneberger O, Fischer P, Brox T. U-net: convolutional networks for biomedical image segmentation [C]//Proc of the 18th International Conference on Medical Image Computing and Computer-Assisted Intervention, Munich: Springer Press, 2015: 234-241.

[21] Gatos B, Ntirogiannis K, Pratikakis I. ICDAR 2009 document image binarization contest (DIBCO 2009) [C]// Proc of the 10th International Conference on Document Analysis and Recognition, Barcelona: IEEE Press, 2009: 1375-1382.

[22] Pratikakis I, Zagoris K, Barlas G, et al. ICFHR2016 handwritten document image binarization contest (H-DIBCO 2016) [C]// Proc of the 15th International Conference on Frontiers in Handwriting Recognition, Shenzhen: IEEE Press, 2016: 619-623.

[23] Pratikakis I, Zagoris K, Barlas G, et al. ICDAR2017 competition on document image binarization (DIBCO 2017) [C]// Proc of the 14th IAPR International Conference on Document Analysis and Recognition, Kyoto: IEEE Press, 2017: 1395-1403.

Note: Figure translations are in progress. See original paper for figures.