Improved Degraded Document images Using Phase Based Binarization

[1]Naga Sudha D, [2]Y Madhavee Latha, [3]L Pratap Reddy

[1]JNTU Jagityal, [2]JNTU Hyderabad, [3]JNTU Hyderabad

Abstract: Image binarization plays major role for document image binarization. Scanning and printing of documents can degrades their visibility that makes difficult to understand them. This paper has proposed a new technique which has ability to binarized documents in efficient manner. In the proposed method, a phase-based binarization model for document images is proposed, as well as a post processing method that can improve any binarization method and a ground truth generation tool. In the pre-processing and binarization steps the features are mainly phase derived while in the post-processing specialized adaptive Gaussian and median filters are considered..This technique also uses adaptive Gaussian filter to improve the accuracy rate further. Finally ground truth generation tool called Phase GT to simplify and speed up the ground truth generation process for document images.

Keywords: Documents, Binarization ,binary image, adaptive Gaussian filter, Ground truth generation.

INTRODUCTION:

Image binarization is the process of separation of pixel values into dual collections, black as foreground and white as background. In degraded documents extensive background noise or difference in contrast and brightness exists.ie there exists many pixels that cannot categorized as for ground or background. To preserve cultural heritage and human knowledge in the form of written texts and printed books, libraries, national archives and projects like Google Books of Google Inc. started a mass digitalization. Additional projects, like Improving Access to Text (IMPACT1) and manuscript research centers (e.g. Vestigia - The Manuscript Research Centre of Graz University2), have the aim to digitize and improve the access to historical documents. The acquired image data needs an automated processing (Document Image Analysis (DIA)) and additionally in the case of historical documents a digital restoration . The research topics of this thesis comprise DIA pre-processing, specifically document binarization, document skew estimation and form classification.

Document pre-processing is the first step of DIA systems and is defined as noise removal and binarization. Additional pre-processing steps of DIA systems are a skew estimation and document classification, e.g. form classification . The skew estimation can be based on binarized images or on original gray value images. Uncorrected documents can effect the performance of Optical Character Recognition (OCR) and segmentation .Document classification can be used for automated indexing in digital libraries by classifying all “Table of Contents” pages or allows a document retrieval on large document image databases . Chen and Blostein state that “document classification is used to tune Optical Character Recognition (OCR) parameters, or to choose an appropriate OCR system for a specific type of document” . By classifying document types a-priori knowledge can be incorporated into the DIA system, thus facilitating higher-level document analysis. While binarization and skew estimation are defined as classical pre-processing steps and form classification

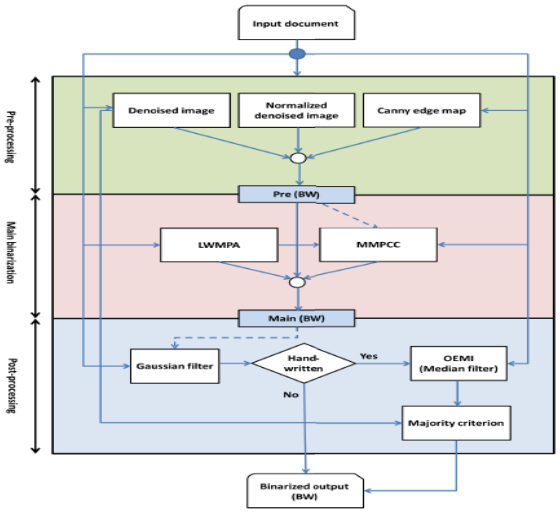


Fig.1 Flowchart of the proposed binarization method

PROPOSED SYSTEM:

A phase-based binarization model for ancient document images is proposed as well as a post-processing method that can improve any binarization method and a ground truth generation tool The proposed model consists of three standard steps 1) pre-processing 2) main binarization and 3) post-processing. In the pre-processing and main binarization steps, the features used are mainly phase derived, while in the post-processing step, specialized adaptive Gaussian and median filters

are considered. One of the outputs of the binarization step, which shows high recall performance, is used in a proposed post-processing method to improve the performance of other binarization methodologies. Finally, we develop a ground truth generation tool, called Phase GT, to simplify and speed up the ground truth generation process for ancient document images. Phase-preserving denoising followed by morphological operations are used to pre-process the input image. The proposed algorithm requires less memory and runs faster. Increase in efficiency compared to other methods.

The de-noising process consists of determining a noise threshold at each scale and shrinking the magnitudes of the filter response vector appropriately, while leaving the phase unchanged. Automatic estimation of these noise thresholds, using the statistics of the smallest filter scale response, is the most important part of de-noising. These statistics are used to estimate the distribution of the noise amplitude, because they give the strongest noise response. Then, the noise amplitude distribution of other filter scales can be estimated proportionally.

Supposing that μR denotes the mean and σ2R denotes the variance of the Rayleigh distribution, the noise shrinkage threshold can be computed using equation. For each orientation, noise responses from the smallest scale filter pair are estimated and a noise threshold is obtained. This noise response distribution is used to estimate the noise amplitude distribution of other filter scales using some constant. Finally, based on the noise thresholds obtained, the magnitudes of the filter response vectors shrink appropriately, and they do so by soft thresholding, while leaving the phase unchanged.

BINARIZATION MODEL:

The final binarized output image is obtained by processing the input image in three steps: pre-processing, main binarization, and post-processing

PREPROCESSING:

In the pre-processing step, we use a denoised image instead of the original image to obtain a binarized image in rough form. The image denoising method is applied to pre-process the binarization output. A number of parameters impact the quality of the denoised output image(ID), the key ones being the noise standard deviation threshold to be rejected (k), and the number of filter scales (Nρ) and the number of orientations (Nr) to be used. The Nρ parameter r controls the extent to which low frequencies are covered. To solve this problem, we combine this binarized image with an edge map obtained using the Canny operator Canny operator is applied on the original document image and for combination those edges without any reference in the aforementioned binarized image are removed. We then compute a convex hull image of the combined image. There are several parameters to be considered in the calculation of IM and IL ..

Main Binarization: The next step is the main binarization, which is based on phase congruency features: i) the maximum moment of phase congruency covariance (IM); and ii) the locally weighted mean phase angle (IL).

1) IM: In this paper, IM is used to separate the back ground from potential foreground parts. This step performs very well, even in badly degraded documents, where it can reject a majority of badly degraded background pixels by means of a noise modeling method. To achieve this, we set the number of two-dimensional log-Gabor filter scales ρ to 2, and use 10 orientations of two-dimensional log-Gabor filters r.

2) IL : We consider the following assumption in classifying foreground and background pixels using IL : where P(x) denotes one image pixel; and IOtsu, bw denotes the binarized image using Otsu's method. Because of the parameters used to obtain the IM and IL maps, IL produces some classification errors on the inner pixels of large foreground objects.

PHASE-DERIVED FEATURES: We use three phase-derived feature maps of the input document image in this paper: two phase congruency feature maps and a denoised image. Phase congruency based feature maps are maximum moment of phase congruency covariance (MMPCC) and local weighted mean phase angle (LWMPA), respectively, at each pixel of input image. Phase congruency first defined terms of a Fourier series expansion of signal.

Maximum moment of phase congruency covariance (MMPCC):Maximum moment of phase congruency covariance (MMPCC)map is a measure of the edges strength. The MMPCC map takes values between[0 1], where a larger value means a stronger edge.

Regional minima: Bleed through degradation is an important and common interfering pattern in the old and historical document images. In this paper, bleed-through is categorized into two classes: i) local bleed-through and ii) global bleed-through. The local bleed-through applies to those degraded pixels where are located under or near the foreground pixels, while the global bleed-through refers to those pixels which are located far from the foreground text. The global bleed-through is one of most challenging degradation because there is no local reference in order to distinguish between the true text and bleed-through. Here, an unsupervised method based on regional minima is proposed to overcome the problem of global bleed-through.

Adaptive Gaussian Filter: In this approach, for each pixel in the image a threshold is selected by calculating local weighted mean along the row, or pairs of rows using a recursive . Afterward, Bradley et al [18] modied this approach by using the integral. However, we used a Gaussian smoothing lter to obtain local weighted mean as the reference value for setting threshold for each pixel

Local weighted mean phase angle**:** Second measure of phase congruency is the local weighted mean phase

angle(LWMPA) at every point in the image which is calculated using equation .The values of this map are between –Π/2 and +Π/2, where a dark line take a value of –Π/2, and a bright line take a value of +Π/2.

Adaptive median filter: It is known that median Filter can reject salt and pepper noises in presence of edges. Similar to the method used in section 4.3 for Gaussian Filter, local thresholds are computed by applying a 4 * 4 symmetric median Filter for each pixel in the input image. In turns, a Filtered image with equal size to the input image is produced medimage, where its pixel values are local thresholds. A pixel is set to 0 (dark) if the value of that pixel in the input image is less than 90% of corresponding pixel value in medical mage, and pixel will set to 1 (white) otherwise.

Phase Preserving De-noising: To be able to preserve the phase data in an image we have to first extract the local phase and amplitude information at each point in the image. This can be done by applying(a discrete implementation of) the continuous wavelet transform and using wavelets that are in symmetric/antisymmetricpairs. Here we follow the approach of Morlet, that is, using wavelets based on complex valued Gabor functions - sine and cosine waves, each modulated by a Gaussian [8]. Using two Filters in quadrature enables one to calculate the amplitude and phase of the signal for a particular scale/frequency at a given spatial location.

Analysis of a signal is done by convolving the signal with each of the quadrature pairs of wavelets. If we let I denote the signal and M_e^n and M_o^n denote the even-symmetric and odd-symmetric wavelets at a scale n we can think of the responses of each quadrature pair of Filters as forming a response vector,

$$[\, e_n(x), o_n(x) \,] = [\, I(x) * M_n^e, I(x) * M_n^o \,] .$$

The values en(x) and on(x) can be thought of as real and imaginary parts of complex valued frequency component. The amplitude of the transform at a given wavelet scale is given by

$$A_n(x) = \sqrt{e_n(x)^2 + o_n(x)^2}$$

and the phase is given by

$$\Phi_n(x) = atan2(o_n(x), e_n(x)).$$

At each point x in a signal we will have an array of these response vectors, one vector for each scale of Filter.

POST PROCESSING: In this step, we apply enhancement processes. First, a bleed through removal process is applied. Then, a Gaussian filter is used to further enhance the binarization output and to separate background from foreground, and an exclusion process is applied, based on a median filter and IM maps, to remove background noise and objects. Finally, a further enhancement process is applied to the denoised image.

Global Bleed-Through Exclusion: Bleed-through degradation is a common interfering pattern and a significant problem in old and historical document images. In this paper, bleed-through is categorized in two classes:

Local bleed-through involves pixels located under or near foreground pixels, while global bleed-through involves pixels located far away the foreground text.

Global bleed-through is one of most challenging forms of degradation, because there is no local to enable true text to be distinguished from bleed-through.

At this stage, we investigate the possibility of the existence of global bleed-through. If it does exist, the parameters of the Canny edge detector are chosen to ensure that the output edge map contains only the edges of text regions which we expect to be located in a specific part, or parts, of the image. The existence of bleed-through is established by comparing the Otsu's result and the binary output obtained so far. If there is a noticeable difference between these two binary images, we apply a global bleed-through exclusion method. Fig. 7 provides two examples of the global bleed-through exclusion process.

Adaptive Gaussian Filter: In this section, we take a similar approach to the one used in , except that a Gaussian smoothing filter is used to obtain a local weighted mean as the reference value for setting the threshold for each pixel. We use a rotationally symmetric Gaussian low-pass filter (G) of size S with σ value, estimated based on average stroke-width, where Fig. 7. Effect of using the proposed global bleed-through exclusion is shown in column (c). The left image (b) is the binarized image before the global bleed-through exclusion step has been applied. σ is the standard deviation. This is a modification of the fixed S value used in [19]. The value for S is the most important parameter in this approach. Local thresholds can be computed using the following two-dimensional correlation:

$$T(x, y) = \sum_{i=-S}^{S} \sum_{j=-S}^{S} G(i, j) \times I(x + i, y + j),$$

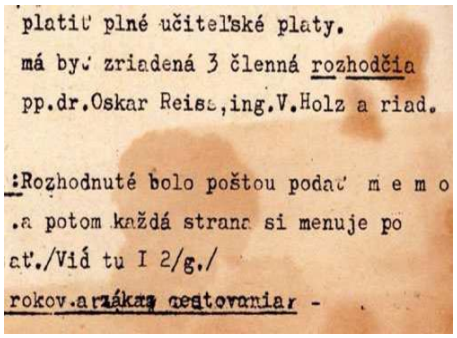
where I (x, y) is a gray-level input image. The result is a filtered image T (x, y) which stores local thresholds. A pixel is set to 0 (dark) if the value of that pixel in the input image is less than 95% of the corresponding threshold value T (x, y),and it is set to 1 (white) otherwise. We increased the value from 85% to 95%, in order to obtain a near optimal recall value.

CONCLUSIONS:

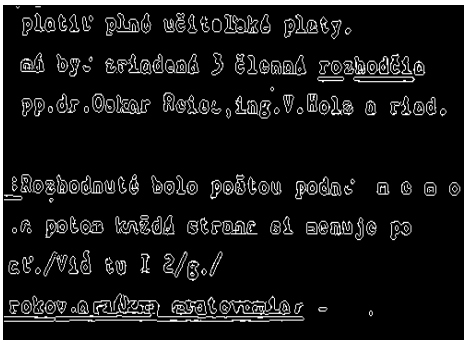
Document binarization is chief application of vision process.The main aim of this paper is to evaluating the shortcomings of the algorithms for degraded image binarization,It has been found that every technique has its own advantages and limitations;no technique is best for each case.The main limitations of existing methods

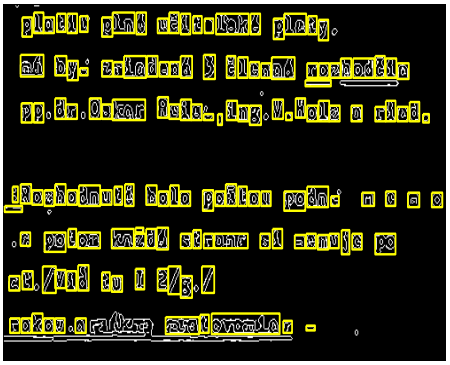
are found that images are noisy and low intensity.The proposed algorithm has used adapative gaussian filter to enhance the results for noisy images.

orginal degraded documnet

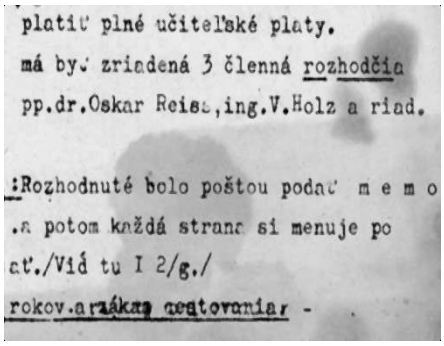


canny edge detection image

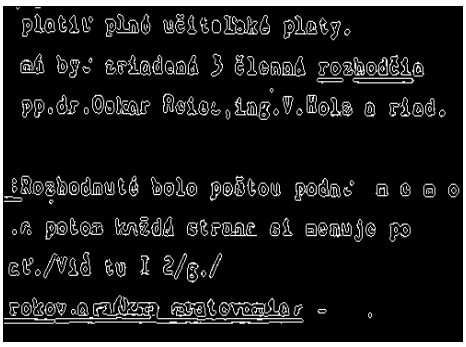


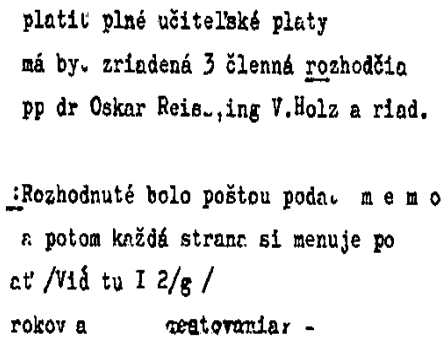


denoised image



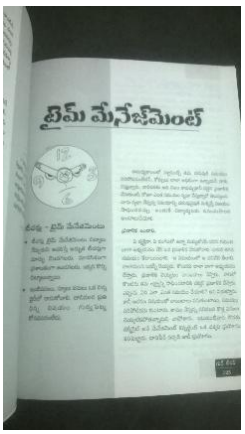
canny edge detection image



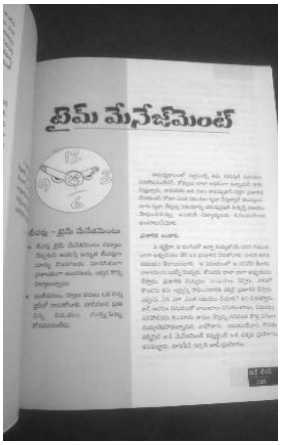


Example of the steps used in the pre-processing phase of the proposed method. a) Denoised image. b) Normalized denoised image. c) Binarization of the original image using Otsu's method. d) Binarization of the normalized denoised image using Otsu's method. e) Edge image using the Canny operator. f) Combination of (d) and (e). g) Convex hull image of (f). h) Combination of images (a) and (g

orginal degraded documnet

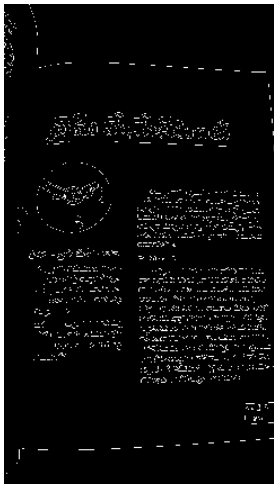


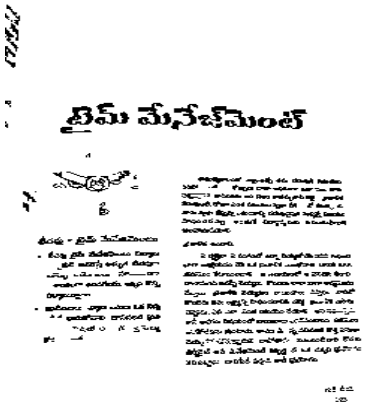
documnet gray image



denoised image

canny edge detection image





REFERENCE:

[1]. Hossein Ziaei Nafchi, Reza Farrahi Moghaddam, Member, IEEE, and Mohamed Cheriet, Senior Member, IEEE. "Phase-Based Binarization of Ancient Document Images: Model and Applications" IEEE TRANSACTIONS ON IMAGE PROCESSING, VOL. 23, NO. 7, JULY 2014

[2]. Florian Kleber and Robert s ablatnig –"Ancient Document analysis Based on Text Line Extraction",2008 IEEE

[3] Spitz,A.L:Determination of the Script and Language Content of Document Images. IEEE Transactions on Pattern Analysis and Machine Intelligence 19:3,235-245(1977)

◈◈◈