ISSN (Online) 2394-6849

International Journal of Engineering Research in Electronics and Communication Engineering (IJERECE)
Vol 4, Issue 3, March 2017

A Novel Ancient Document Image Binarization Technique

[1]P.Gangadhara Reddy, [2] Dr.T.Ramashri
[1] Research scholar ,[2] Professor
SVUniversity College of Engineering,Tirupathi,A.P.,INDIA

Abstract: **-- Due to many environmental factors, improper handling and the poor quality of the materials used in the creation of old document images cause them to suffer a high degree of degradation which includes faded ink, bleed-through, show-through, uneven illumination, variations in image contrast and deterioration of the cellulose structure. In proposed algorithm a robust phase-based binarization technique is used for the binarization and enhancement of historical documents and manuscripts. The main object of proposed method consists of preprocessing, main binarization, and post-processing. The preprocessing step mainly involves image denoising with phase preservation, followed by some morphological operations to denoise the image and detection of edge information by using canny edge detector. Then, the phase congruency features used for the main binarization step. The foreground of ancient documents can be modeled by phase congruency. After completing the three binarization steps on the input images using phase congruency features and a denoised image, the enhancement processes are applied. A median filter and a phase congruency feature are used to construct an object exclusion map image. This map is then used to remove unwanted lines and interfering patterns. 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.**

Keywords: **- Historical document binarization, Phase congruency feature, Document enhancement.**

I. INTRODUCTION

An adaptive binarization method based on low-pass filtering, foreground estimation, background surface computation, and a combination of these. A binarization method based mainly on background estimation and stroke width estimation. First, the background of the document is estimated by means of a one-dimensional iterative Gaussian smoothing procedure. Then, for accurate binarization of strokes and sub-strokes, an L1 -norm gradient image is used.
The local maximum and minimum is used to build a local contrast image. Then, a sliding window is applied across that image to determine local thresholds. Learning-based methods have also been proposed in recent years. These methods are an attempt to improve the outputs of other binarization methods based on a feature map, or by determining the optimal parameters of binarization methods for each image. The disadvantages of the above method are

- The existing system cannot deal with different sort of ancient documents and different types of degradations.
- Less efficiency since it produces only rough binarization.

Motivation

Libraries and archives around the world store an abundance of old and historically important documents and manuscripts. These documents accumulate a significant amount of human heritage overtime. However, many environmental factors, improper handling, and the poor quality of the materials used in their creation cause them to suffer a high degree of degradation of various types. Today, there is a strong move toward digitization of these manuscripts to preserve their content for future generations. The huge amount of digital data produced requires automatic processing, enhancement and recognition. A key step in all document image processing workflows is binarization, but this is not a very sophisticated process, which is unfortunate, as its performance has, a significant influence on the quality of OCR results.

Problem Definition

Due to many environmental factors, improper handling and the poor quality of the materials used in the creation of old document images cause them to suffer a high degree of degradation which includes faded ink, bleed-through, show-through, uneven illumination, variations in image contrast and deterioration of the cellulose structure.
There are also differences in patterns of hand-written and machine-printed documents, which add to the difficulties associated with the binarization of old document images.
None of the proposed methods can deal with all types of documents and degradation. In this paper, a robust phase-based binarization method is proposed for the binarization and enhancement of historical documents and manuscripts.

ISSN (Online) 2394-6849

International Journal of Engineering Research in Electronics and Communication Engineering (IJERECE)
Vol 4, Issue 3, March 2017

The three main steps in the proposed method are: preprocessing, main binarization, and post-processing. The preprocessing step mainly involves image denoising with phase preservation, followed by some morphological operations. We incorporate the canny edge detector and a denoised image to obtain a binarized image in rough form.
Then, the phase congruency features used for the main binarization step. Phase congruency is widely used in the machine vision and image processing literature, palmprint verification, object detection, finger-knuckle print recognition, and biomedical applications are just a few examples of the use of phase congruency as a feature detector. The foreground of ancient documents can be modeled by phase congruency. After completing the three binarization steps on the input images using phase congruency features and a denoised image, the enhancement processes are applied. A median filter and a phase congruency feature are used to construct an object exclusion map image. This map is then used to remove unwanted lines and interfering patterns. The effect of each step on the binarized output image is discussed in each associated section.

The web images in the inter.net are often susceptible to certain image degradation such as low resolution and small size, which is specially designed for faster network transmission rate, computer-generated-character artifacts, and special effects on images to attract visual attention.
Document Image Enhancement is a technique that improves the quality of a document image to enhance human perception and facilitate subsequent automated image processing. It is widely used in the pre-processing stage of different document analysis tasks. Document image enhancement problem is essentially an ill-posed problem, because a number of enhanced images can be generated from the same input image. Moreover, the quality of enhancement techniques is mainly judged human perception, which makes the quantitative measures hard to be applied.
The notations used in this paper are given below:

I_L	Local weighted mean phase angle(LWMPA)
I_M	Maximum moment of phase congruency covariance(MMPCC)
$I_{Otsu,bw}$	Otsu's output when applied on I
I_{Pre}	Output of preprocessing step

II.PROPOSED SYSTEM

Introduction
A phase-based binarization model for ancient document images is proposed as well as a post processing method that can improve any binarization method.
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, Phase-preserving denoising followed by morphological operations are used to preprocess the input image.
System Architecture: The proposed method follows the system architecture shown in figure 1.

Preprocessing
In the preprocessing step, a denoised image used instead of the original image to obtain a binarized image in rough form. The image denoising method is applied to preprocess the binarization output. A number of parameters impact the quality of the denoised output image (I_D), the key ones being the noise standard deviation threshold to be rejected (k), and the number of filter scales (N_P) and the number of orientations (N_r) to be used.
We combine the 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.
At the end of this step, the structure of foreground and text is determined. However, the image is still noisy, and the strokes and sub-strokes have not been accurately binarized. Also, the binarization output is affected by some types of degradation.

ISSN (Online) 2394-6849

International Journal of Engineering Research in Electronics and Communication Engineering (IJERECE)
Vol 4, Issue 3, March 2017

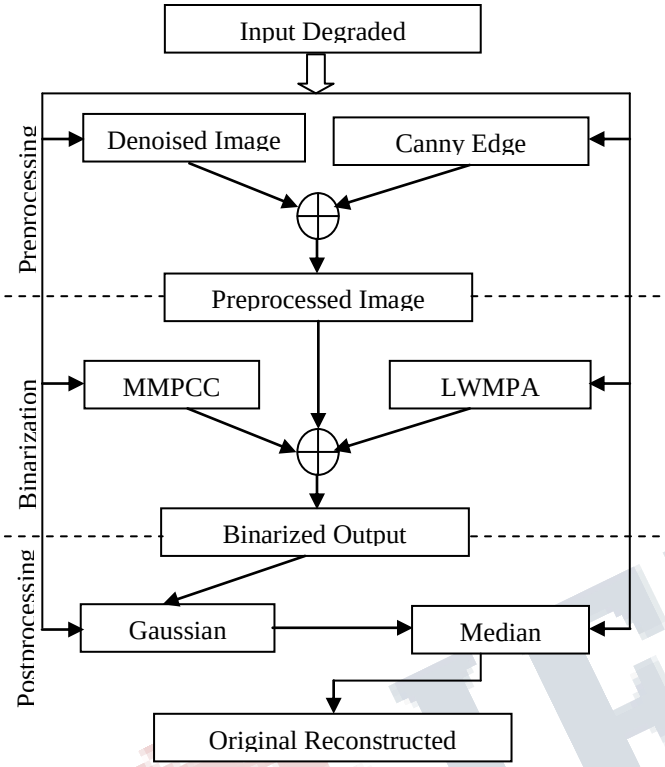


Fig.1 : System Architecture

Phase preserving denoising of images

Denoising of images is typically done with the following process: The image is transformed into some domain where the noise component is more easily identified, a thresholding 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/antisymmetric pairs. 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. 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. However, rather than using Gabor filters we prefer to use *log Gabor* functions as suggested by Field ; these are filters having a Gaussian transfer function when viewed on the logarithmic operation is then applied to remove the noise, and finally the transformation is inverted to reconstruct a (hopefully) noise-free image.

To preserve the phase data in an image have to first extract the local phase and amplitude information at each frequency scale. Log Gabor filters allow arbitrarily large bandwidth filters to be constructed while still maintaining a zero DC component in the even-symmetric filter. A zero DC value cannot be maintained in Gabor functions for bandwidths over 1 octave. It is of interest to note that the spatial extent of log Gabor filters appears to be minimized when they are constructed with a bandwidth of approximately two octaves. This would appear to be optimal for denoising as this will minimize the spatial spread of wavelet response to signal features, and hence concentrate as much signal energy as possible into a limited number of coefficients.

Example of preprocessing:

Fig(a) ***fig(b)***

Fig. 2: Example of the steps used in the pre-processing phase of the proposed method. A) Denoised image. b) Normalized denoised image.

Main Binarization

The next step is the main binarization, which is based on phase congruency features: i) the maximum moment of phase congruency covariance (I_M); and ii) the locally weighted mean phase angle (I_L). There has been a renewed interest in the detection of so called `corners', or `interest points'. The success of these reconstructions depends very much on the reliable and accurate detection of these points across image sequences. The definition of a corner is typically taken to be a location in the image where the local autocorrelation function has a distinct peak.

Another difficulty many of these operators have is that the Gaussian smoothing that is employed to reduce the influence of noise can corrupt the location of corners, sometimes considerably. The SUSAN operator deserves some special comment here because it does not suffer from these problems outlined above.

It identifies features by determining what fraction of a circular mask has values the same, or similar, to the value at the centre point. Thresholds are therefore defined in terms of the size of the mask and no image smoothing is required.

ISSN (Online) 2394-6849

International Journal of Engineering Research in Electronics and Communication Engineering (IJERECE)
Vol 4, Issue 3, March 2017

However, the SUSAN operator assumes that edges and corners are formed by the junctions of regions having constant, or near constant, intensity, and this limits the junction types that can be modeled.

To address the many problems outlined in this paper describes a new corner and edge detector developed from the phase congruency model of feature detection. The new operator uses the principal moments of the phase congruency information to determine corner and edge information. Phase congruency is a dimensionless quantity and provides information that is invariant to image contrast. This allows the magnitudes of the principal moments of phase congruency to be used directly to determine the edge and corner strength. The minimum and maximum moments provide feature information in their own right; one does not have to look at their ratios. If the maximum moment of phase congruency at a point is large then that point should be marked as an edge. If the minimum moment of phase congruency is also large then that point should also be marked as a 'corner'. The hypothesis being that a large minimum moment of phase congruency indicates there is significant phase congruency in more than one orientation, making it a corner.

The resulting corner and edge operator is highly localized and the invariance of the response to image contrast results in reliable feature detection under varying illumination conditions with fixed thresholds. An additional feature of the operator is that the corner map is a strict subset of the edge map. This facilitates the cooperative use of corner and edge information. This is organized as follows: first the phase congruency model of feature perception is reviewed. We then examine how the phase congruency responses over several orientations can be analyzed in terms of moments to provide both edge and corner information. Finally the performance is assessed relative to the commonly used Harris operator.

1) **$\mathbf{I_M}$** *:* In this paper, I_M is used to separate the background 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 . In addition, the number of standard deviations k used to reject noises is estimated as follows:

$$k = 2 + (\alpha)\left[\frac{\sum_{n,m} \mathrm{IOtsu,bw}\,(n,m)}{\sum_{n,m} \mathrm{Ipre}(n,m)}\right] \quad \ldots(1)$$

where α is a constant (we are using α = 0.5); $I_{\mathrm{Otsu,bw}}$ is the binarization result of Otsu's method on the input image; and I_{pre} is the output of the preprocessing step. Here, the minimum possible value for k is 2. Note the different values used for setting the phase congruency feature and denoised image parameters. We use I_M to remove a majority of the background pixels.

2) **$\mathbf{I_L}$** *:* We consider the following assumption in classifying foreground and background pixels using I_L

$$P(x) = \begin{cases} 1, IL(x) < 0 \\ 0, IL(x) > 0 \,\&\, \mathrm{IOtsu,bw} = 0 \end{cases} \quad \ldots(2)$$

where $P(x)$ denotes one image pixel; and $I_{\mathrm{Otsu,bw}}$ denotes the binarized image using Otsu's method. Because of the parameters used to obtain the I_M and I_L maps, I_L produces some classification errors on the inner pixels of large foreground objects. Using more filter scales would solve this problem, but reduce the performance of I_L on the strokes. Also, I_L impacts the quality of the I_M edge map, and of course requires more computational time. Nevertheless, the results of using Otsu's method to binarize the large foreground objects are of interest.Consequently, we used the $I_{\mathrm{Otsu,bw}}$ image t0 overcome the problem.

Example of main binarization:

Fig(a) ***fig(b)***

ISSN (Online) 2394-6849

International Journal of Engineering Research in Electronics and Communication Engineering (IJERECE)
Vol 4, Issue 3, March 2017

Fig(c) *fig(d)*

Fig(e) *fig(f)*

Fig. 3: A degraded document image and its binarized image using phase -congruency. a) Original degraded document image. b) Edge image obtained by phase congruency (I_M). c) Filled image of I_M . d) Binarization of (c) using Otsu's method. e) Denoised image and f) The result of main binarization

Image Binarization using Local Maximum and Minimum:

This presents a simple but efficient historical document image binarization technique that is tolerant to different types of document degradation such as uneven illumination and document smear. The proposed technique makes use of the image contrast that is evaluated based on the local maximum and minimum. Given a document image, it first constructs contrast image and then extracts the high contrast image pixels by using Otsu's global thresholding method.

After that, the text pixels are classified based on the local threshold that is estimated from the detected high contrast image pixels. The proposed method has been tested on the dataset that is used in the recent DIBCO contest series. Experiments show that the proposed method outperforms most reported document binarization methods.

Post Processing

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

Global Bleed-Through Exclusion

Bleed-through degradation is a common interfering pattern and a significant problem in old and historical document images. In this project, bleed-through is categorized in two classes: i) local bleed- through; and ii) global bleed-through. 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.

Adaptive Gaussian Filter

In this section, 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 σ is the standard deviation. This is a modification of the fixed S value . The value for S is the most important parameter in this approach. Local thresholds can be computed using the following two-dimensional correlation:

$$T(x,y) = \sum_{i=-S}^{S} \sum_{j=-S}^{S} G(i,j) * I(x+i, y+j) \quad \ldots(3)$$

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% [47] to 95%, in order to obtain a near optimal recall value.

ISSN (Online) 2394-6849

International Journal of Engineering Research in Electronics and Communication Engineering (IJERECE)
Vol 4, Issue 3, March 2017

Some sub steps of the proposed binarization method work on objects rather than on individual pixels, and so it is important to separate foreground text from the background. The Gaussian filter described above is one of the methods used to achieve this. This filter is also applied to the equalized adaptive histogram image instead of the original image, in order to preserve weak strokes.

The average stroke width is computed, in order to set *S*. There are various methods for computing Stroke width. In this paper, a very rapid approach, based on the reverse Euclidean distance transformation of the rough binary image obtained so far, is used to estimate the average stroke width. This approach is dependent on the quality of the rough binary image, which has the potential to produce errors; however, it is a very fast way to calculate stroke width, and provides a good estimate of the average stroke width.

Document type detection

At this step, we need to determine the type of input document we are dealing with. We propose to apply the enhancement processes that are after this step to the handwritten documents only, and not to machine- printed documents. The method we propose for detecting the type of document is straightforward and fast. We use the standard deviation of the orientation image that was produced during calculation of the phase congruency features.

This image takes positive anticlockwise values between 0 and 180.A value of 90 corresponds to a horizontal edge, and a value of 0 indicates a vertical edge. By considering the foreground pixels of the output binary image obtained so far, we see that the standard deviation value of the orientations for these pixels is low for handwritten document images and higher for machine-printed documents.

The reason for this is the different orientation values for interior pixels and edges. This approach works well for almost all the images we tested, including 21 machine-printed images and 60 handwritten document images, and only one classification error was found. In below figure we can see the histogram of orientations of a handwritten document follows a U-shape behavior. Note that even if this approach fails to accurately detect the type of document, it nevertheless produces satisfactory output.

Median Filter

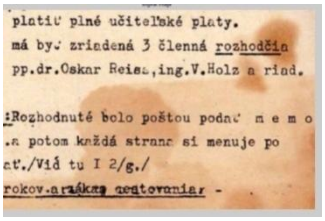
Median filtering is a nonlinear method used to remove noise from images. It is widely used as it is very effective at removing noise while preserving edges. It is particularly effective at removing 'salt and pepper' type noise. The median filter works by moving through the image pixel by pixel, replacing each value with the median value of neighboring pixels. The pattern of neighbors is called the "window", which slides, pixel by pixel over the entire image 2 pixel, image. The median is calculated by first sorting all the pixel values from the window into numerical order, and then replacing the pixel being considered with the middle (median) pixel value.

Median filtering is one kind of smoothing technique, as linear Gaussian filtering. All smoothing techniques are effective at removing noise in smooth patches or smooth regions of a signal, but adversely affect edges. Often though, at the same time as reducing the noise in a signal, it is important to preserve the edges. Edges are of critical importance to the visual appearance of images, for example. For small to moderate levels of (Gaussian) noise, the median filter is demonstrably better than Gaussian blur at removing noise whilst preserving edges for a given, fixed window size. However, its performance is not that much better than Gaussian blur for high levels of noise, whereas, for speckle noise and salt and pepper noise (impulsive noise), it is particularly effective. Because of this, median filtering is very widely used in image processing.

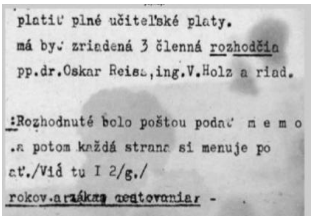
Algorithm description of median filter

The main idea of the median filter is to run through the signal entry by entry, replacing each entry with the median of neighboring entries. The pattern of neighbors is called the "window "which slides, entry by entry, over the entire signal. For 1D signals, the most obvious window is just the first few preceding and following entries, whereas for 2D (or higher-dimensional) signals such as images, more complex window patterns are possible (such as "box" or "cross" patterns). Note that if the window has an odd number of entries, then the median is simple to define: it is just the middle value after all the entries in the window are sorted numerically. For an even number of entries, there is more than one possible median.

III.RESULTS & ANALYSIS



fig(a)



fig(b)

ISSN (Online) 2394-6849

International Journal of Engineering Research in Electronics and Communication Engineering (IJERECE)
Vol 4, Issue 3, March 2017

fig(c) *fig(d)*

fig(e) *fig(f)*

fig(g) *fig(h)*

fig(i) *fig(j)*

fig(k) *fig(l)*

fig(m) *fig(n)*

Fig 4: a) Original Degraded Document b)Gray Image c)Denoised Image d)Edge Image e) Preprocessed Image f)MMPCC Image g)LWMPA Image h)Combined Phase Image i)Binarized Image j) Gaussian Filtered Image1 k)Gaussian Filtered Image2 l)Median Filtered Image1 m)Median Filtered Image 2 n) Original Reconstructed Image

IV. CONCLUSION:

In this paper, an image binarization method that uses the phase information of the input image, and robust phase-based features extracted from that image are used to build a model for the binarization of ancient manuscripts. Phase-preserving denoising is used to preprocess the input image. Then, two phase congruency features, the maximum moment of phase congruency covariance and the locally weighted mean phase angle, are used to perform the main binarization. For post-processing, we have proposed a few steps to filter various types of degradation, in particular, a median filter has been used to reject noise, unwanted lines, and interfering patterns. Because some binarization steps work with individual objects instead of pixels, a Gaussian filter was used to further separate foreground from background objects, and to improve the final binary output. Our experimental results demonstrate its promising performance, and also that of the postprocessing method proposed to improve other binarization algorithms.

REFERENCES

[1] Hossein Ziaei Nafchi, Reza Farrahi Moghaddam, and Mohamed Cheriet,"Phase-based binarization of ancient document images: Model and applications"in IEEE Transactions on Image Processing,2014

[2] P. Kovesi, "Phase preserving denoising of images," in *Proc. Int. Conf. Digital Image Comput., Techn. Appl.*, 1999.

[3] J. Canny, "A computational approach to edge detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 8, no. 6, pp. 679–698, Nov. 1986.

[4] P. Kovesi, "Image features from phase congruency," *Videre, J. Comput.Vis. Res.*, vol. 1, no. 3, pp. 1–26, 1999.

[5] H. Z. Nafchi and H. R. Kanan, "A phase congruency based document binarization in *Proc. IAPR Int. Conf. Image Signal Process.*, 2012, pp. 113–121.

[6] H. Z. Nafchi, R. F. Moghaddam, and M. Cheriet, "Historical document binarization based on phase information of images," in *Proc. ACCV*, 2012, pp. 1–12.