ISSN 2250-3153

Phase-Based Binarization of Ancient Document Images

Ms. Priyanka R. Patil*, **Prof. Dr. S. L. Lahudkar****

* Electronics & Tele-comm. engineering department, JSPM's Imperial College of Engineering And Research, Pune.
** Electronics & Tele-comm. engineering department, JSPM's Imperial College of Engineering And Research, Pune.

Abstract- The main defects present in historical documents are darkness, non-uniform clarification, bleed-through and faded characters. To remove these defects binarization method is used. In this paper a phase based binarization method is studied in which phase of ancient document images is preserved. This method is derived in to three steps: preprocessing, main binarization and post processing. In preprocessing phase preserved denoised image is derived. In main binarization two phase feature maps are derived are maximum moment of phase congruency covariance and a locally weighted mean phase angle. At last in post processing Gaussian and median filter is use for enhancement of image. It is also improve the performance of binarization methodologies.

Index Terms- Historical document binarization, phase-derived features, document enhancement.

I. Introduction

HISTORICAL documents go through different degradations due to getting old, complete use, some attempts of acquisition and ecological situation. The main defects present in historical documents are darkness, non-uniform clarification, smear, strain, bleed-through and faded characters. Those defects are problematic for document image analysis methods which presume even background and consistent quality of writing. In handwritten documents, the writer may use different amount of ink and pressure and make characters of dissimilar intensity or thickness, as well as faint characters. The same writer may write in different ways even within the same document. Similar problems, such as faint characters and non-uniform appearance of characters of the same font, are also encountered in historical machine-printed documents.

Today, there is a strong progress in the direction of digitization of these historical documents to save their content for future generations. The huge quantity of digital data created requires automatic processing, enhancement, and recognition. A key step in all document image processing workflows is binarization, but this is not a very complicated process, which is unfortunate, as its performance has a significant control on the quality of OCR results. lots of research studies includes the problem arising due to binarization of old document images characterized by many types of degradation and they are finding solution, including faded ink, bleed-through, show-through, uneven illumination, variations in image contrast. There are also variations in patterns of hand-written and machine-printed documents, which add to the difficulties related with the binarization of old document images.

A phase-based binarization method is proposed for the binarization and enhancement of historical documents and manuscripts. The three main steps in the proposed method

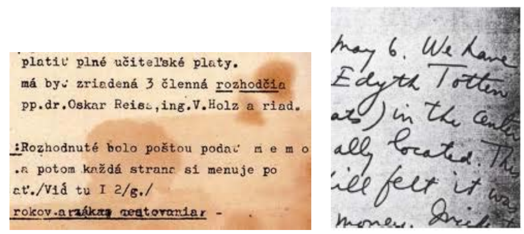

Fig. 1. Sample document images

A phase-based binarization method is proposed for the binarization and enhancement of historical documents and manuscripts. The three main steps in the proposed method are: preprocessing, main binarization, and post-processing. Phase of the image does not corrupt so phase information is mainly preserved. The preprocessing step involves image denoising. In which phase information is mainly preserved, followed by some morphological operations. Then edge map image is obtained by using canny edge operator. These two images are combined to obtained binarization image in rough form.

Then, for the main binarization step we use the phase congruency features. Phase congruency is dimensionless quantity that is invariant to changes in image brightness and contrast. 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.

.

II. LITERATURE REVIEV

In this section, some binarization methods are briefly described. Gatos et al. [4] propose an adaptive binarization method based on low-pass filtering, foreground estimation, background surface computation, and a combination of these. In [5], an initial binary map is obtained using the multi-scale Sauvola's method , and then statistical methods are used to restore the missed strokes and sub-strokes. In [7], Valizadeh et al. map input images into a two-dimensional feature space in which the foreground and background regions can be distinguished. Then, they partition this feature space into several

ISSN 2250-3153

small regions, which are classified into text and background based on the results of applying Niblack's method.

Lu et al. propose 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. This method placed 1st of 43 algorithms submitted to the DIBCO'09 competition. In [1], a local contrast image is combined with a Canny edge map to produce a more robust feature map.

Farrahi Moghaddam et al. [1] propose a multi-scale binarization method in which the input document is binarized several times using different scales. Then, these output images are combined to form the final output image. This method uses different parameters for Sauvola's method to produce output images of the same size, but at different scales. In contrast, Lazzara and Gerard propose a multi-scale Sauvola's method which binarizes different scales of the input image with the same binarization parameters. Then, binary images with different scales are combined in some way to produce the final results.

Combination methods have also attracted a great deal of interest, and provided promising results. The goal of combining existing methods is to improve the output based on assumption that different methods complement one another. In [10], several of these methods are combined based on a vote on the outputs of each. In [6], a combination of global and local adaptive binarization methods applied on an inpianted image is used to binarize handwritten document images. The results show that this method performs extremely well; however, it is limited to binarizing handwritten document images only.

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 [8]–[9], or by determining the optimal parameters of binarization methods for each image In [8] and [9], a self-training document binarization method is proposed. The input pixels, depending on the binarization method(s) used are divided into three categories: foreground, background, and uncertain, based on a priori knowledge about the behavior of every method used. Then, foreground and background pixels are clustered into different classes using the k-means algorithm or the random Markov field [8], [9]. Finally, uncertain pixels are classified with the label of their nearest neighboring cluster. The features used for the final decision are pixel intensity and local image contrast.

Another combined method based on a modified contrast feature is proposed. Lelore and Bouchara also classify pixels into three categories using a coarse thresholding method, where uncertain pixels are classified based on super resolution of likelihood of foreground. Howe proposes a method to optimize the global energy function based on a Laplacian image. In this method, a set of training images is used for optimization. Howe improved this method by tuning two key parameters for each image. In [10], a learning framework is proposed to automatically determine the optimal parameters of any binarization method for each document image. After extracting the features and determining the optimal parameters, the relation between the features and the optimal parameters is learned. As we show in the Experimental Results and Discussion section, a problem associated with all these algorithms is that they are not reliable for all types of degradation and with different datasets.

III. METHODOLOGY

The final binarized output image is obtained by processing the input image in three steps: preprocessing, main binarization, and postprocessing.

1. Preprocessing

In the preprocessing step, denoised image instead of the original image is used to obtain a binarized image in rough form. 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 ($N\rho$) and the number of orientations (Nr) to be used. The $N\rho$ parameter controls the extent to which low frequencies are covered. The higher $N\rho$ is, the lower the frequencies, which means that the recall value remains optimal or near optimal. Based on our experiments, $N\rho = 5$ is the appropriate choice in this case. Therefore, to preserve all the foreground pixels, we set the parameters in the experiments as follows: $k = 1$, $N\rho = 5$ and $Nr = 3$.

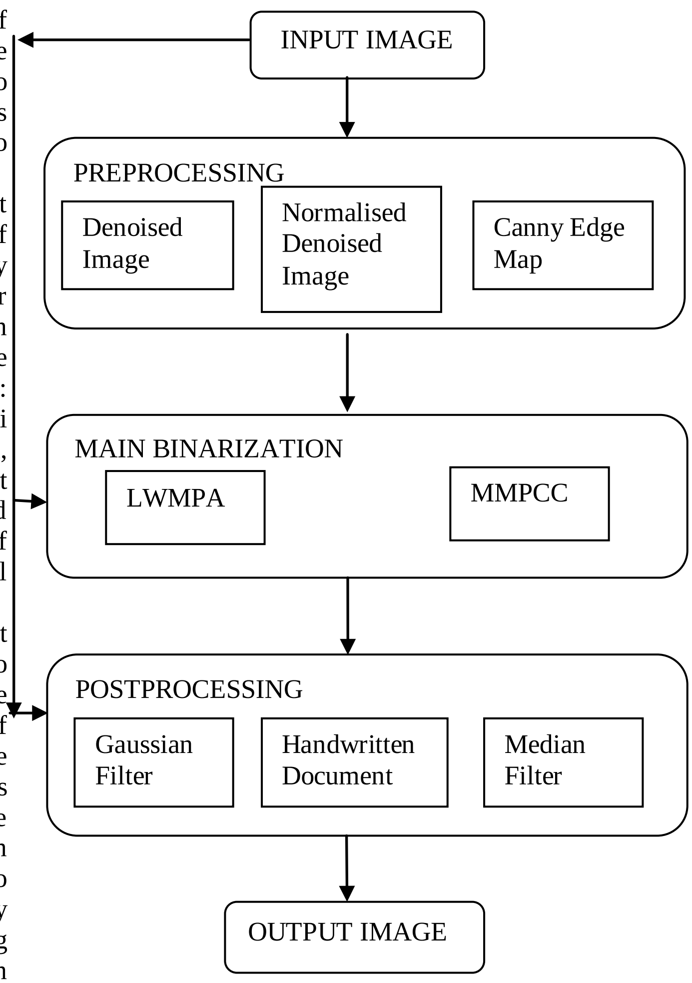

Fig.2. Flowchart of the proposed binarization method.

We used Otsu's method on the normalized denoised image, where normalized denoised image is obtained by applying a linear image transform on the denoised image. This approach can also remove noisy and degraded parts of images, because the denoising method attempts to shrink the amplitude information of the noise component. The problem with this approach is that it misses weak strokes and sub-strokes, which means that we cannot rely on its output. 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.

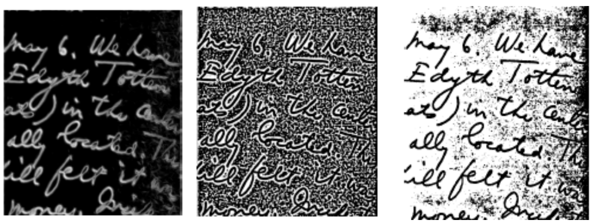

Fig.3. Example of *IM*, *IL*, and *ID* maps.

2. 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*).

a) IM*:*

In this paper, *IM* 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. Two-dimensional phase congruency is calculated by

$$PC_{2D,r}(x) = \frac{\sum_p W_r(x)[A_{pr}(x)\Delta\Phi_{pr}(x) - T_r]}{\sum_p A_{pr}(x)}$$

Two-dimensional phase congruency is calculated by:

$$I_M = \max_r PC_{2D,r}(x)$$

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:

$$k = 2 + \left[a \times \left(\frac{\sum_{n,m} I_{Otsu,bw}(n,m)}{\sum_{n,m} I_{pre}(n,m)}\right)\right]$$

(20)

where α is a constant (we are using $\alpha = 0.5$); *I*Otsu,*bw* isthe 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.

5.2.2 IL:

The two-dimensional locally weighted mean phase angle (*IL*) is obtained using the summation of all filter responses over all possible orientations and scales:

$$I_L(x) = arctan\left[\sum_{p,r} e_{pr}(x), \sum_{p,r} o_{pr}(x)\right]$$

We consider the following assumption in classifying foreground and background pixels using *IL* :

$$p(x) = \begin{cases} 1, & I_L(x) \le 0 \\ 0, & I_L(x) > 0 \; and \; I_{Otsu,bw}(x) = 0 \end{cases}$$

(21)

where *P(x)* denotes one image pixel; and *I*Otsu,*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. Using more filter scales would solve this problem, but reduce the performance of *IL* on the strokes. Also, *IL* impacts the quality of the *IM* 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*Otsu,*bw* image to overcome the problem.

3.Postprocessing:

In this step, we apply enhancement processes. First, a bleedthrough 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. The individual steps are as follows.

IV. CONCLUSION

The information of phase for input image using the image binarization method and extraction of robust phase-based features from that image are needed to build a model for the binarization of ancient documents. Depends the morphological operations preprocess the input image. Then, main binarization is performing the phase congruency features. In post-processing, a proposed gaussian and trimmed median filter has been needed for eliminate noise, unwanted lines, and interfering patterns. A Gaussian filter was helpful for separating the both foreground, background objects, and also improving the final binary output. The manual correction is deduced based on this tool which is involved in ground truth generation.The application of phase-derived features,the stable behavior of document images, to other cultural heritage fields,can be maintained for long time.So, they are very useful to our future generation to follow the ancient culture and their traditions. And also our historical documents can be saved.

REFERENCES

[1] B. Su, S. Lu, and C. L. Tan, "Robust document image binarization technique for degraded document images," IEEE Trans. Image Process., vol. 22, no. 4, pp. 1408–1417, Apr. 2013.

[2] R. F. Moghaddam and M. Cheriet, "AdOtsu: An adaptive and parameterless generalization of Otsu's method for document image binarization," Pattern Recognit., vol. 45, no. 6, pp. 2419–2431, 2012.

[3] J. Sauvola and M. Pietikinen, "Adaptive document image binarization," Pattern Recognit., vol. 33, no. 2, pp. 225–236, 2000.

[4] B. Gatos, I. Pratikakis, and S. Perantonis, "Adaptive degraded document image binarization," Pattern Recognit., vol. 39, no. 3, pp. 317–327, 2006.

[5] R. Hedjam, R. F. Moghaddam, and M. Cheriet, "A spatially adaptive statistical method for the binarization of historical manuscripts and degraded document images," Pattern Recognit., vol. 44, no. 9, pp. 2184–2196, 2011.

[6] K. Ntirogiannis, B. Gatos, and I. Pratikakis, "A combined approach for the binarization of handwritten document images," Pattern Recognit. Lett., vol. 35, pp. 3–15, Jan. 2014.

[7] B. Su, S. Lu, and C. Tan, "Binarization of historical document images using the local maximum and minimum," in Proc. 9th IAPR Int. Workshop DAS, 2010, pp. 159–166.

[8] B. Su, S. Lu, and C. L. Tan, "A self-training learning document binarization framework," in Proc. 20th ICPR, Aug. 2010, pp. 3187–3190.

[9] B. Su, S. Lu, and C. L. Tan, "A learning framework for degraded document image binarization using Markov random field," in Proc. 21st ICPR, Nov. 2012, pp. 3200–3203.

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

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

[12] K. Ntirogiannis, B. Gatos, and I. Pratikakis, "A performance evaluation methodology for historical document image binarization," IEEE Trans. Image Process., vol. 22, no. 2, pp. 595–609, Feb. 2013.

AUTHORS

First Author – Ms. Priyanka R. Patil, Electronics & Tele-comm. engineering department, JSPM's Imperial College of Engineering And Research, Pune, E-mail ID- priyankarpatil65@gmail.com

Second Author – Prof. Dr. S. L. Lahudkar, Electronics & Tele-comm. engineering department, JSPM's Imperial College of Engineering And Research, Pune, E-mail ID- swapnillahudkar@gmail.com