International Journal of Scientific Research & Engineering Trends
Volume 3, Issue 1, Jan.-2017, ISSN (Online): 2395-566X

An Assessment of Handle Noise in Digitization of Historical Document

Anand Singh Rajput*, Saurabh Pandey, Anubhav Sharma
Acropolis Institute of Technology & Research, Bhopal
*Email: anandsinghrajput940@gmail.com

***Abstract* – The document analysis systems often begin at an early stage binarization processing. Although many techniques have been proposed binarization, the results can vary in quality and are often sensitive to the parameters of one or more control parameters. This paper discusses a promising approach for the binarization based on simple principles, and shows that its success depends on the most important values of two key parameters. A method to automatically adjust the settings so that the air in the individual image is also described, giving a final binarization algorithm that can reduce the total error of a third compared to the basic version. The results of this method to advance the state of art in the last reference point.**

***Keywords* – DIBCO, Document Binerization, GEF, MRF, PSNR, MSE.**

I. INTRODUCTION

Physical documents in the real world have a multitude of colors and shades, and the appearance of a single document can vary greatly depending on factors such as lighting, viewing angle, etc digital reproductions of the physical documents are generally using a representation of 24-bit color, or may be 8 bits of grayscale. These formats do not include some of the data present in the original, but retain more than enough for most applications. Some systems document processing goes much further: they retain a single bit per pixel document. Documenting binary representations are useful despite their low fidelity, since information is lost in more or less related to the symbolic content of the document binarization. Most documents are produced using a monochrome ink on paper, and their meanings are incorporated exclusively for the distribution of ink, a bit pattern representing the document explicitly.

Of course, the deduction of correct binarization of a document from color or grayscale representation can be difficult. The physical deterioration of the document, or image illumination conditions and limits of the unfavorable resolution can contribute to obscure the original pattern.1 .To meet this challenge, researchers have proposed many algorithms for document image binarization. In fact, an intense activity in this area was the competition standard document image binarization (Dibco), one of the most popular in the analysis of the research community documents competitions, attracting 17 entries in its last iteration [19]. However, the results of these competitions show that there is always room for improvement in the quality of automatic binarization.

This document makes two contributions to the study of automatic paper binarization. First, a method that can achieve excellent results in a wide variety of images described documents difficult. Secondly, as a means of these results, we studied a method to automatically determine the best values for each particular image parameters. Estimation of the automatic settings has received little research attention to date, but it has two advantages when done right: it makes the easiest method to use because the user does not have to worry about finding the best set of parameters and also improves the average score for each image you can use your own ideal value instead of a single undertaking and running, but sub-optimal for all images.

II. RELATED WORK

Otsu's method is a binarization threshold parameter less overall .It method assumes the presence of two distributions (one for the text and one for the background), and calculates a threshold value in such a distance from minimize the difference between two distributions [30, 45]. The limit for the distribution of two Otsu method was eliminated in [45], where the modes of degradation in the image histogram does one remove one applying recursively Otsu's method until only one mode remains in the picture .In other work, the overall limitation of the method is removed [10] and an adaptation method is introduced, which uses the same concept as the Otsu method, but local patches. A measure based on the global threshold Otsu in this paper to reveal the regions of text are not to have one class of pixels was used. As mentioned in the introduction, this method still requires full implicit minimum (see section 2.11 for details) .In this paper, a new adaptive method using estimated identify regions are not plaintext is proposed. Thus, the method is well suited to the input image. Then two methods are discussed locally adapted thresholding.

Among the binarization threshold scalable approaches, the Sauvola [34] method is one of the best known .In this method, the threshold value, inspired by the Niblack method [29,41] has been modified to capture open areas not text is [34]. The threshold has two parameters to define and estimate.

A binarization method state-of-the-art is introduced in [15]. In this method, one first obtains a coarse binarization of the document image (usually using the method Sauvola) a rough background (see section 2.12 below) .Then

estimated. In the next step, local threshold values are calculated based on the estimated background and some parameters .These threshold values are used to calculate the final binarization that is post-processed to remove noise.

The binarization method proposed by Su, Lu and Tan, and first place in the competition DIBCO'09 binarization [14]. The method comprises four steps:

(i) Removing the background by polynomial fit of the lines;

(ii) Detecting the contours of the race using Otsu's method on the gradient information;

(iii) Local threshold by averaging the detected pixels in a local neighborhood window edges, and finally

(iv) Post-processing of results.

The second method was proposed by DIBCO'09 Fabrizio and Marcotegui [14]. It is based on morphological mapping operator shaft rocker [7]. To avoid salt and pepper noise associated with balancing mapping, are excluded from the analysis pixel erosion and dilation are too close .Pixels are then classified as text, background pixels and uncertain .The are uncertain assigned to the text and background, depending on their class boundary.

The method was third in DIBCO'09 was proposed by Rivest-He 'Nault, Farrahi Moghaddam and Cheriet [33,14]. This method uses the level within the limits established to locate text of shots and binarization of the image of a document. [33] Like the others, this algorithm involves several steps:

(I) Initialization using a race card (SM) [9];

(II) The correction of the SM mode with level adjustment frame and eroding local linear models, and finally,

(iii) A second series of level adjustment operations, this time with a gray level of the career force, which provides the final text regions as the inner regions of the overall function of the levels.

Although images of the documents may suffer from degradation severe and vari- measure, we can assume that there are areas that could be described as actual text or background. This hypothesis has been the basis of many learning methods, which are based on a rough estimate of the text boxes and background and then try to learn their behavior to classify regions found in the confusion range [6.39, 37, 3,12,7,20]. For example, a simple threshold to identify the kinds of text and background in [6] was used. Then a noise model is constructed and is used to adjust the threshold value .In [39], a frame using a binarization method is presented to identify three classes; ie text, background, and uncertain .then pixels, the pixels of uncertainty reclassified using a classifier trained using the classes of the text and background. We refer to this framework as a self-learning environment in Table 3.

Local maximum and minimum method

In this method, [38], the contrast of the image, defined on the basis of minimum and local maximum, is used to detect the pixels of a higher contrast image instead of the image contrast gradient is .Image less sensitive to uneven illumination .Then, the document is segmented at a threshold based on the local image contrast estimate.

This method [32], which is an extension of the axis of the component based on the flat areas of hyper links, the axis defines a special command in hyper node connections and enables non-flat areas not .The method as follows: background removal using a hyper tree component; the adaptive threshold values based on image edges are detected by Sobel operator with Otsu thresholding, and finally post-processing.

Thanks to a grid-based modeling introduced in [10], the method of cost calculation Sauvola can be significantly reduced. This allows the introduction of the scale on the basis of a multi methodin Sauvola door [10], which is capable of capturing the text pixel scales and high scales Monthe smaller track patterns to avoid strongly interfere .In work , used a similar through multiple devices with the method AdOtsu approach to improve performance.

III. BINARIZATION

The binarization basic approach used in this work was recently presented [11] and is based on three mutually reinforcing strategies. First, define the target binarization and labeling of pixels that minimizes global energy function based on a design of MRF. Second, in the formulation of the length of data precision of this energy is based on the Laplacian of the image intensity to distinguish background ink. This essential invariance attributed to differences in contrast and overall intensity. Third, it incorporates advanced discontinuities in terms of regularity of the overall function of the power, distorting ink limits to align the edges and allow harder smoothing incentive other image. The following paragraphs describe each of these points in greater detail below.

The global energy function works binarizations B, label each pixel indexed by (i,j) is as ink or background, Bij $\in$ {0, 1}. Energy comes as a typical additive, with terms that reflect the specific labeling fidelity compared to the current data, and other terms representing the smoothness or regularity of the solution. In particular, the energy includes a L0 or L1 cost ij ij understanding how the Bi j label chosen for each pixel corresponds to its appearance, and irregular costs Chij and Cvij for each pixel whose label respectively differ from its neighbor horizontal or vertical.

$$\varepsilon I(B) = \sum_{i=0}^{m}\sum_{j=0}^{n}[L0ij(1-Bij) + L1ijBij] + \sum_{i=0}^{m-1}\sum_{j=0}^{n} C\hbar ij(Bij \neq Bi+1,j) + \sum_{i=0}^{m}\sum_{j=0}^{n-1} C_{vij}\,(Bij \neq Bi,j+1) \ldots\ldots(1)$$

Suppose that the above expressions for evaluating the Boolean 0 or 1 in the usual manner according to the truth value. With this energy, the optimum binarization tend to conform to the contours of intensity while smoothing unevenness resulting from noise sources. The degree of smoothness with respect to the accuracy of the information depends on Lbij relative magnitudes to Chij and Cvij.

The label L0 and L1ij costs should be invariant to the illumination of the work area, and therefore are trapped in the Laplacian of the image intensity:

$$L0ij = \nabla 2Iij$$
$$L1ij = -\nabla 2Iij$$

Intuitively, this tends to separate the ink from the bottom due to the divergence of the gradient of Laplacian measurement. Therefore, it will be positive intensity valleys (ink) and negative current peaks or plateaus (bottom). The data terms of the energy function becomes a sum of the signed label Laplacian in each pixel of the image. For a component or ink particular fund, Green's theorem tells us that the sum of the Laplacian all pixels is mathematically equivalent to the gradient flow across the border. In other words, the energy contribution of each component is determined solely by what happens on its border. This makes intuitive sense, but can cause problems for the components that cut the edges of the image: these areas are sometimes mislabeled because its all natural border is not visible, and therefore the real contribution of the energy cannot be estimated . In practice, this causes problems occasionally for background noise zones are isolated from the rest of the collection of documents with ink marks. There are several possible solutions. For example, we could simply set L1 ij a large negative value for all pixels (i, j) to the edge of the image, assuming that the ink is framed by the background regions. Instead of participating in a strong assumption such, this work has a more conservative strategy, looking for outliers brightest pixels and applying a constant L1ij attached to them. This also ensures that large background regions receive the label itself, and does not prevent the identification of the ink on the pixels of the final image. To be more specific, L1ij to change all pixels of more than two standard deviations σrij brighter than the average μrij in the area, the surrounding pixels calculated by a weighted Gaussian radius r. This can be regarded as a local adaptive thresholding extremely conservative, where only background pixels most are sure to be labeled as such. In the equation, φ has a large negative value.

$$L1ij = \begin{cases} -\nabla 2Iij & Iij \le \mu ij + 2\sigma rij \\ \varphi & Iij > \mu ij + 2\sigma rij \end{cases}$$

The mismatch penalties neighbor Chij and Cvij offer the possibility of using the third strategy mentioned above, incorporating the map Canny (remember that Canny [6] first smooth the image with a Gaussian filter σE small radio, then finds maximum edges directed local gradient of the image, choosing to keep only selected by a process of two thresholds and hysteresis thi and tlo). The algorithm defines the gap penalties to a uniform value c everywhere except between pixels where Canny identified a probable discontinuity. To be more specific, Canny pixels identified as edges, while equation. 1 requires positioning discontinuities at the connections between pairs of pixels. To address this discrepancy, the formulation below zero at the discontinuity penalty of Canny edge pixels and neighbors brighter, effective choice to include Canny pixels in the inked surface. The opposite choice would also make self-consistent, but would prevent the detection of broad strokes of a single pixel.

$$Chij = \begin{cases} 0 & if\ Eij \cap (Iij < Ii+1, j) \\ 0 & if\ Ei+1, j \cap (Iij \ge Ii+1, j) \\ c & otherwise \end{cases}$$

$$Cvij = \begin{cases} 0 & Eij \cap Iij < Ii, j+1) \\ 0 & Ei, j+1 \cap (Iij \ge Ii, j+1) \\ c & otherwise \end{cases}$$

The determination of sanctions discontinuity constant everywhere except at the edges deserves note. One could imagine using a fine that varies continuously depending on the intensity similarity among neighbors. Empirically, this approach seems less effective, perhaps because it actually gives little guidance on the best location precise ink background transition intensity differences tend to be large everywhere after a few pixels the border real, so that it becomes too easy to choose the place.

IV. CONCLUSION

Parameter settings offer both the potential risks and benefits. Selection of parameter values that are optimal for a given image binarization can reduce the error in large quantities and a framework for static generic images.On Moreover, the values of the poorly chosen parameters can sabotage the result: loss potential overall quality binarization overshadow potential gains. So be careful to give only when there is reasonable certainty of success.

This paper presents a heuristic stability criterion to choose the values of the appropriate parameters for each image. The approach assumes that the appropriate values of the parameters are marked by small variations in the binarization solution with respect to changes in the values of the parameters. This approach leads to an algorithm that successfully takes good values of c in all the analyzed images. Similar heuristics applied to the choice of thi also selects good values of the parameters in most cases, although there have been some shortcomings with this technique for a handful of cases. The most successful method tested using the stability criterion to choose c, and thi selected from a limited set of two possible values. This approach provides a significant fraction of the maximum possible gain setting two parameters, and can be calculated at about one-eighth of the speed of the simple binarization

with parameters. Reference static code for the art will be available on the website of author.

The configuration results of the survey, it is evident that many images, the control algorithms given here are about maximizing the potential of the algorithm based on binarization. Other improvements in the quality of the results are likely to come through the development of new algorithms based. Some of these new algorithms can be application-specific, while this paper has sought wide applicability. The stability criterion that proved so useful in this case apply to other approaches remains to be seen, and is a topic for future work. In all cases, the control parameters to the algorithms described here significantly advance the state of the art document binarization, as evidenced by the comparative results of the test images DIBCO 2011.

REFERENCES

[1] Reza Farrahi Moghaddamn , Mohamed Cheriet "AdOtsu: An adaptive and parameterless generalization of Otsu's method for document image binarization" in Elsevier transaction of Pattern Recognition pg no- 2419–2431,2012

[2] B. Gatos, K. Ntirogiannis, I. Pratikakis, ICDAR 2009 document image binarization contest (DIBCO 2009), in: ICDAR'09, 2009, pp. 1375–1382.

[3] Pratikakis, I., Gatos, B., Ntirogiannis, K.: ICDAR 2011 document image binarization contest (DIBCO 2011). In: International Conference on Document Analysis and Recognition, pp. 1506–1510 (2011)

[4] M. Sezgin, B. Sankur, Survey over image thresholding techniques and quantitative performance evaluation, Journal of Electronic Imaging 13 (1) (2004) 146–168.

[5] R. Farrahi Moghaddam, M. Cheriet, A multi-scale framework for adaptive binarization of degraded document images, Pattern Recognition 43 (6) (2010)2186–2198.

[6] B. Gatos, I. Pratikakis, S.J. Perantonis, Adaptive degraded document image binarization, Pattern Recognition 39 (3) (2006) 317–327.

[7] B. Gatos, K. Ntirogiannis, I. Pratikakis, Dibco 2009: document image binarization contest, International Journal on Document Analysis and Recognition (2010) 1–10.

[8] J. Fabrizio, B. Marcotegui, M. Cord, Text segmentation in natural scenes using toggle-mapping, in: ICIP'09, 2009, pp. 2373–2376.

[9] B. Gatos, K. Ntirogiannis, I. Pratikakis, ICDAR 2009 document image binarization contest (DIBCO 2009), in: ICDAR'09, 2009, pp. 1375–1382.

[10] R. Hedjam, R. Farrahi Moghaddam, M. Cheriet, A spatially adaptive statistical method for the binarization of historical manuscripts and degraded document images, Pattern Recognition 44 (9) (2011) 2184–2196.

[11] B. Su, S. Lu, C.L. Tan, A self-training learning document binarization frame work, in: ICPR'10, 2010, pp. 3187–3190.