Indian Journal of Science and Technology, *Vol 10(20), DOI: 10.17485/ijst/2017/v10i20/73915, May 2017*
ISSN (Print) : 0974-6846
ISSN (Online) : 0974-5645

Analysis of Document Images through Local Binary Patterns

N. Anusha[1]* and M. Siva Sindhuri[2]

[1]Computer Science and Engineering, Sathyabama University, Chennai – 600100, Tamil Nadu, India; anusha.nallapareddy@gmail.com
[2]Computer Science and Engineering, Vaagdevi Institute of Technology and Science, Proddatur – 516360, Andhra Pradesh, India

Abstract

Objective: To develop a robust Local Binary Pattern (LBP) method with ENI (Edge pixels, Noisy pixels and Interior pixels) features for enhancement and binarization of ancient document images. **Methods:** The Local Binary Pattern method is used for improving the image quality by the extraction optimal image features. This method works efficiently for noisy and low contrasted ancient document image. The foreground and background of the low contrasted ancient document images are overlapped and in some documents back ground may be very complex. It is very difficult to separate background from foreground of the document Images. These flaws can be removed effectively by robust Local Binary Pattern method. This method can separate the foreground and background of the ancient document images by enhancing the quality. **Findings:** Thus the proposed LBP method achieves better results even when more level of noise is added to the image**.** It enhances and extracts the text in the document image by giving the robust results. The experimental results show that the proposed method produces better results than traditional methods. **Applications:** The proposed Local Binary Pattern method can be used to improve and extract the optimal image features from the ancient or low contrasted and noisy document images.

Keywords: Ancient Documents, Binarization, Image Classification, Local Binary Pattern (LBP)

1. Introduction

Today, many historical documents are stored in the archives and libraries across the world. But all those documents are not clear due to the usage of low quality material in those days. All those historical documents can be viewed with the help of digitization technique which is emerged in the present era. In the binarization of very aged document images many challenges are faced. So in order to clear those problems many research studies have also been carried. In the present days, the main research studies have been carrying on image classification. But even though extracting an image features for noisy image is still a great challenge.

In the existing method of binarization, Kovesi's method is used. The binarization is performed mainly through three phases. They are:

1. Preprocessing
2. Main Binarization
3. Post processing

Consider g(x) as 1 dimensional signal for a pair of filters at an image point x.

$$[f_p(x), o_p(x)]=[g(x)*Nfp, g(x)*N^o_p]$$

Now they found the local phase $Ø_p(x)$ as follows:

$$Ø_p(x)=\arctan2(o_p(x), f_p(x))$$

$$A_p(x)=\text{sqrt}(f_p(x)^2+o_p(x)^2)$$

*Author for correspondence

$$\Delta\emptyset_p=\cos(\emptyset_p(x)-\acute{\emptyset}(x))-|\sin(\emptyset_p(x)-\acute{\emptyset}(x))|$$

Where,

$\Delta\emptyset_p$ denotes a sensitive phase deviation function.

To find the noise threshold the following form is used:

$T=\mu R+K\boldsymbol{\sigma}_R$

But this Kovesi's method is very high sensitive to noise.

Pre-processing stage involves the process of getting binarized image in rough form. In the pre-processing step of the binarization method, degraded ancient document image is considered and morphological operations on the image are applied which produces a de-noised image. After obtaining the de-noised image, the canny edge detector is applied in order to get binarized image in rough form. Then in the main binarization which is nothing but the conversion of image to only black and white, phase congruency features are used in order to construct an object exclusion map image. This object exclusion map image is used to remove unwanted lines.

In the post processing of binarization method, enhancement process is applied. Here first, a bleed through removal process is applied and then the foreground and background is separated using Gaussian filter. Based on this filter it removes noise and objects.[1] In case of the binarization method, the main problems are it is less efficient for low contrasted images and noisy images and it is not robust in nature.

In the reported article it proposed a new heuristic binarization approach in order to deal with the complex backgrounds.[1] In the next reported article described the problems associated with quality assessment and comparison of thematic maps that are generated from Remote Sensing (RS) images.[2] It proposed an original data driven thematic map quality assessment. The Remote sensing images are generally used for identification of map errors. The article proposed a novel document binarization method.[3] It addresses the issues by using adaptive image contrast which is the couple of local image gradient. This method is adaptive to text and variation of background and local image contrast. The proposed method is a tested on public datasets which are used in the latest document image binarization.

In the article conducted a survey on the existing method of binarization and evaluation measurement.[4] This leads to the invention of a new method for the automatic selection of the binarization method in order to handle historical document images. The main use of binarization method is to convert the grey scale image to binary image. In the reported article presented a Histogram of Oriented Gradient based two-directional Dynamic Time Warping matching method for handwritten word spotting.[5] In article it proposed a technique called Semantic annotated Markovian Semantic Indexing which is used for retrieving the images and automatically annotates the non-annotated images in the database using hidden Markov model.[6] In the reported article presented a new framework for text detection in historical handwritten document images, created a new dataset containing medieval document images and marked ground truth text areas.[7] In the article described how to use a computer algorithm for automatically quantify bone morphology.[8] And in this paper we use very important tool called 3-dimensional X-ray micro tomography in order to investigate bone morphology in and human biopsy here many researchers have been investigated some standard methods for determining a grey level threshold value.

In the article the results of a research project are found by applying various types of geospatial information technologies.[9] Here it becomes very complex to use the old methods in order to conduct archaeological studies, which is mainly based on the technology of remote sensing. In the described article a technique called phase based binarization and ground truth generation tool is represented.[10] These two techniques are mainly used for the enhancement of old documents and the manuscripts. The existing system mainly comprises of three stages .They are preprocessing, main binarization and post processing. The first two stages are mainly used for phase derived and the third stage done using some specialized filters. It is not fully efficient for low contrasted document images and it is not robustness.

In the reported article a novel strategy for constructing steganography detectors for digital images is described.[11] Here we propose a common methodology for steganalysis of digital images which is generally based on the concept of a rich model comprising of a more number of diverse sub models. In the reported article a solution to the binarization problem of technical document images is given.[12] The technical document images are generally used for the construction of plants, ships or vehicles which is the prerequisite for efficient digital storage and transmission. In the article a method called novel adaptive binarization is used which is based on wavelet filter.[13] This method shows good performance on complex images and it process faster compare to the other adaptive methods. In this paper for the eval-

uation of binarization method we use a method called goal-directed. In the mentioned article described mainly two methods in which the first method is depends on an algorithm called multidimensional scaling.[14] Whereas the second method uses an approach called graph based. Here documents are generally represented by the vertices of a graph. But here in both of these methods it permits only a limited amount of data.

In the article extracted the text from the historical document images with the usage of a method called robust segmentation.[15] This robust segmentation method is mainly depends on Markovian-Bayesian clustering. This method can able to preserve the edges and weak connections which are very necessary in the skeletonization steps that are mainly depends on topological features of shapes.

In the reported article it developed a technique called fundus camera imaging which is necessary to examine the retina in order to detect the various disorders of the Age- related macular degeneration, diabetic retinopathy and glaucoma.[16] This paper describes a method for retinal image set. This method also presents a tool called validation for tracking back the distortion path.

In the reported article it presented a technique called novel document image binarization.[17] It is used to separate the information from background. Here the methods of binarization is classified into two types i.e., global and local methods. In proposed article adaptive binarization algorithm which shows best results even in the case where the input images are distorted highly.[18] This paper also found the adaptive processing on various types of images. It also describes the methods for thresholding images in order to produce binary images.

In the article described how to extract the local image features for classification of images.[19] This paper introduces a Local Binary Pattern method with the help of ENI (Edge pixels, Noisy pixels and Interior pixels). This method is very simple and powerful for analyzing the images. But it failed to detect large-scale image structures. In the described article a large quantity of data is necessary in order to estimate the parameters of a high dimensional model.[20] So it uses two methods for describing large quantity of data. The first method is re-ranking approach and the second method is automatic query generation in which high number of queries is generated automatically for parameter estimation. In the described article efficient binarization algorithm with the help of intelligent block size detection.[21] Binarization is nothing but the process of conversion of grey level images into binary images. Here the method called Otsu is proposed. But it failed to obtain the expected output in the case where the source image contains different background patterns (or) inhomogeneous background.

In the article described how to detect the localize texts in natural scene images.[22] Text detection is very important for content-based image analysis. Here the Connected Component (CC) based and Region based methods have several problems. In Region based method the performance is very sensitive to text alignment orientation and speed is also very low. In CC based method it can't segment text components accurately. In the described article the multi-resolution image representation by applying 1D and 2D filters.[23] In this paper the proposed a combination of 1D separable and 2D no separable filter banks in order to reduce the articrafts.Eventhough this method shows good results nonlinear approximation redundancy is not desirable in several kinds of practical signal processing. The reported article used threshold method to locally update the threshold value.[24] Here binarization is mainly performed with the help of thresholding which is broadly classified into two classes called global and local. But these two classes will not be sufficient for the poor quality of the source document.

2. Proposed System

We propose Local Binary Pattern (LBP) method which mainly introduces three types of pixels. They are:

1. Noisy pixels
2. Edge pixels
3. Interior pixels

Figure 1(a), (b), (c), shows the local neighborhood all these 3 types of pixels denote a number of homogeneous pixels.

10	19	12	0	11
16	11	2	44	66
21	19	-	56	31
57	11	11	22	55
29	1	10	21	11

Figure 1(a). T_{5x5} q=22.

12	3	34	22	11
6	33	20	66	88
1	3	22	78	53
79	33	11	44	77
51	21	32	43	11

Figure 1(b). E(q,r)=|T(q)-T(r)|.

0	0	1	1	0
0	1	1	1	1
0	0	-	1	1
1	1	0	1	1
1	0	1	1	0

Figure 1(c). Z=15.

2.1 Algorithm

Step 1: To find the difference between central pixel q and its neighborhood pixel r.

$$E(q,r)=|T(q)-T(r)|$$

Step 2: From step1 the gray level values are obtained. Made those values into two groups.

where,

p is a gray value of central pixel.

Step 3: To find the total no of object pixels we use the following formula

$$Z=\Sigma\ n\text{-}1i=0t(r)$$

Step 4: Find rotation invariant by using the following formula

$$M=1/4\text{-}1/4(\cos(\ \Pi Z\ /\ N)\ )$$

where,

N is the number of pixels in T and Z.

Step 5: Find the edge preserving value by

$$A_B=(T\ (f,g)\text{-}T(f,g\text{-}1))/2$$
$$A_{BB}=(T(f,g+1)+T(f,g\text{-}1))\text{-}2^{*}T(f,g)$$
$$A_C=(T(f+1,g)\text{-}T(f,g)/2$$
$$A_D=(T(f+1,g)+T(f\text{-}1,g))\text{-}2^{*}T(f,g)$$
$$S=T(f,g)+(A_C{}^{2*}A_{BB}(2^{*}A_B{}^{*}A_C)+A_B{}^{2*}A_D)/(A_B+A_C)$$

where,

A_B,A_C are the first order gradients along with the directions A and C.

A_{BB},A_D are the second order gradients and S represents the edges in noisy images.

3. Experimental Results and Discussions

The performance of the LBP method has been tested by considering a document image shown. The proposed LBP method produced very good results than the existing binarization method. The existing binarization method is very high sensitive to noise. Whereas the LBP method showed better results for both low contrasted and noisy images. This LBP method is very efficient and robust in nature.

The salt and pepper noise as shown in Figures 2(a),2(b),2(c),2(d),2(e),2(f) & 2(g) is added in increasing order (see Column 1). The existing method failed to remove noise in the document image at the 30% of noise (see Column 2),whereas the proposed method (see Column 3) was able to produce better results up to 50% of noise and it failed to remove noise in the document image at 60% of noise. So by this, we can say that the proposed method can yield better results than the existing method.

Figure 2(a). Noiseless.

Figure 2(b). Addition of 10% noise.

Figure 2(c). Addition of 20% noise.

Figure 2(d). Addition of 30% noise.

Figure 2(e). Addition of 40% noise.

Figure 2(f). Addition of 50% noise.

Figure 2(g). Addition of 60% noise.

Figures 2(a), 2(b), 2(c), 2(d), 2(e), 2(f) and 2(g).
Column 1 shows Results of document images after adding salt and pepper noise.
Column 2 Results of extracting document images by using existing method and
Column 3 Results of extracting document images by using proposed method.

The Mean (MN) and the Standard Deviation (SD) values for both existing method and proposed method by adding the salt and pepper noise are given in the increasing order from noiseless to 60% of added noise are given. Figure 3 the Mapping between existing and proposed mean values and Figure 4 shows the mapping between existing and proposed standard deviation values.

4. Conclusion

The new technique called Local Binary Pattern's (LBP) is proposed, which can be used for extracting data from degraded images. The older binarization methods failed to extract the data from images when there is 30% of noise,

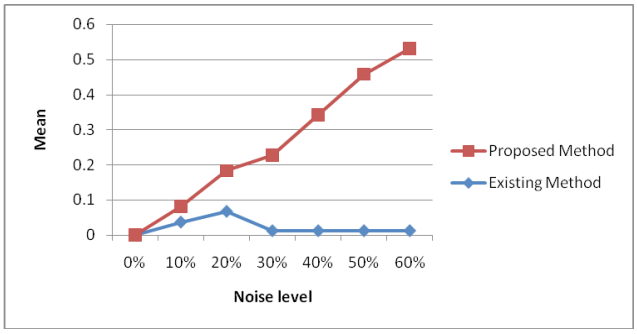

Figure 3. Mapping between Existing Method mean and Proposed Method mean.

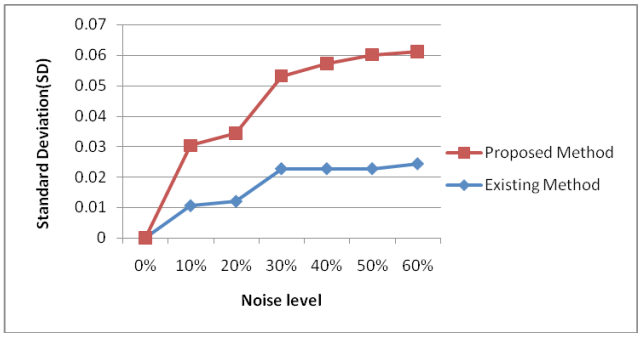

Figure 4. Mapping between Existing Method Standard Deviation (SD) and Proposed Method Standard Deviation (SD).

whereas the proposed method was able to enhance the image quality of ancient document images, extract the data from images and able produce reliable results even at 50% of noise. In future the same proposed technique can be further developed to extract the data from the image, even it is more degraded.

4. References

1. Baraldi A, Bruzzone L, Blonda P. Quality assessment of classification and cluster maps without ground truth knowledge. IEEE Transactions on Geosciences and Remote Sensing. 2005 Apr; 43(4):857–73. Crossref
2. Su B, Lu S, Tan CL. Robust image binarization technique for degraded document images. IEEE Transactions on Image Processing. 2013 Apr; 22(4):1408–17. Crossref PMid:23221822
3. Fung CC, Chamchong R. A review of evaluation of optimal binarization technique for character segmentation in historical manuscripts. Third International Conference on Knowledge Discovery and Data-mining. Phuket Thailand; 2010 Mar. p. 236–40. PMid:20478804

4. Sales E, Gomez W, Pereira WCA. Evaluation performance of local adaptive binarization algorithms Trabecular Bone on simulated UCT. IEEE Nuclear Science Symposium and Medical Imaging Conference. Valencia Convention Center Valencia Spain; 2012 Feb. p. 3084–7.
5. Yao S, Wen Y, Lu Y. HoG based two-directional Dynamic Time Warping for handwritten word spotting. 13th International conference on Document Analysis and Recognition; 2015 Nov. p. 161–5. Crossref
6. Sangeetha M, Anandakumar K. Annotation based Image Retrieval System by Mining of Semantically Related User Queries with Improved Markovian Model. Indian Journal of Science and Technology. 2015 Jun; 8(35):1–7. Crossref
7. Hofmann S, Gropp M, Bernecker D, Pollin C, Maier A, Christlein V. Vesselness for text detection in historical document images. IEEE Conference Publications; 2016 Aug. p. 3259–63. Crossref
8. Mao F, Li Q, Huang J, Tang Z, Zhoo W. The application of Remote sensing technology in the archaeological study of the segment of grand canal in Shandong province. IEEE International Geoscience & Remote Sensing Symposium (IGARSS). John B. Hynes Veterans Memorial Convention Center Boston Massachusetts USA; 2009 Feb. p. 48–51.
9. Cavalcanti GDC, Silva EFA, Zanchettin C, Bezerra BLD, Doria RC, Rabelo JCB. A Heuristic Binarization Algorithm for Documents with Complex Background. IEEE International Conference on Image Processing. Atlanta GA USA; 2007 Feb. p. 389–92.
10. Nafchi HZ, Moghaddam RF, Cheriet M. Phase-based binarization of ancient document images Model and Applications. IEEE Transactions on Image Processing. 2014 Jul; 23(7):2916–30. Crossref PMid:24816587
11. Fridrich J, Kodavsky J. Rich Models for Steganalysis of Digital Images. IEEE Transactions on Information Forensics and Security. 2012 Jun; 7(3):868–82. Crossref
12. Valverde JS, Grigat RR. Optimum binarization of technical document images. International Conference on Image Processing; 2002 Aug. p. 985–8.
13. Yang J, Jiaofeng K, Jiao L, Xu JJ. A fast adaptive binarization method for complex scene images. 19th IEEE International Conference on Image Processing; 2013 Feb. p. 1889–92.
14. Thiel K, Dill F, Kotter T, Berthold MR. Towards visual Exploration of Topic Shifts. IEEE International Conference on SMC; 2008 Jan. p. 522–7.
15. Hedjam R, Moghaddam R F, Cheriel M. Text extraction from degraded document images. 2nd European workshop on Visual Information Processing; 2011 Jan. p. 247–52.
16. Lee S, Abramoff MD, Reinhardt JM. Validation of Retinal image registration algorithms by a projective imaging distortion model. 29th Annual International Conference of the IEEE EMBS; 2007 Oct. p. 6471–4.
17. Tabatabaei SA, Bohlool M. A novel method for binarization of badly illuminated document image. IEEE 17th International Conference on Image Processing; 2010 Dec. p. 3573–6. Crossref
18. Mo S, Mathew VJ. Adaptive quadratic preprocessing of document images for binarization. IEEE Transactions on Image Processing. 1998 Jul; 7(7):992–9. Crossref PMid:18276315
19. Suresh B, Salma S, Reshma A, Nagaraju C. Local Binary Patterns with ENI features for images classification and analysis. I-Managers Journal on Electronics Engineering. 2011 Sep; 2(1):45–50.
20. Oba T, Hori T, Nakamura A, Ito A. Spoken document retrieval by discriminative modeling in a high dimensional feature space. IEEE International Conference on Acoustic Speech and Signal Processing (ICASSP); 2012 Aug. p. 5153–6. Crossref
21. Chang YF, Pai YT, Ruan SJ. An efficient thresholding algorithm for degraded document images based on intelligent block detection. IEEE International Conference on Systems Man and Cybernetics; 2009 Apr. p. 667–72.
22. Pan YF, Hou X, Liu CF. A Hybrid Approach to Detect and Localize Texts in Natural Scene Images. 15th IEEE Transactions on Image Processing. 2011 Mar; 20(3):800–13. Crossref PMid:20813645
23. Tanaka Y, Ikehara M, Nguyen TQ. A new combination of 1D and 2D filter banks for effective multi resolution image representation. IEEE International Conference on Image Processing; 2008 Dec. p. 2820–3. Crossref
24. Rangasanseri Y, Rodtook S. Comparative study of thresholding techniques for gray-level Document Image Binarization. IEEE 10th International conference on Electrical and Electronic Technology; 2002 Aug. p. 152–5.