ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 3, March 2015

Document Image Binarization Using Threshold Segmentation

Rekha Chaudhari[1], Dinesh Patil[2]

Student, Dept. of CSE, SSGBCOET, Jalgaon, Maharashtra, India[1]

Associate Professor and HOD, Dept. of CSE. SSGBCOET Jalgaon, Maharashtra, India[2]

ABSTRACT: Binarization is process to generate binary image from document image. Document image binarization has already under research from past many years, and many binarization algorithms have been proposed for different types of degraded document images. Document image Binarization is very popular to upgrade old handwritten and machine printed documents. Still to recover degraded document is very tedious job. Such document has the much damaged also presence of noise and degradation. There is a lot of scope to improve old and degraded documents. Image segmentation is method which used frequently in image processing. Thresholding is an important pre-processing step for the degraded image to enhance their quality. The between the foreground text and the background of different document images is a difficult task. New Binarization method using image segmentation using threshold segmentation is proposed. Proposed method can overcome the drawback of canny edge map.

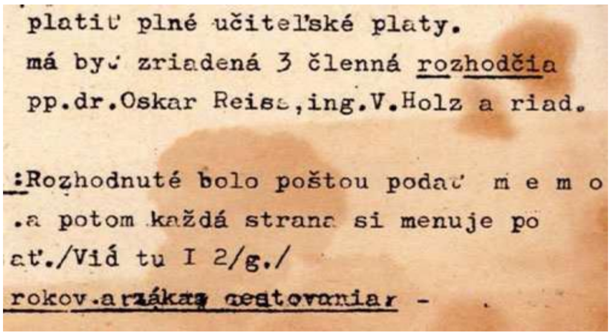
KEYWORDS: Image Segmentation, Threshold segmentation, Document image Binarization, Thresholding.

I. Introduction

Document image binarization is used to convert document image into gray scale image and then into binary image. Document image binarization is important step in document image processing and analysis. Document analysis involves handwriting recognition, optical character recognition (OCR), extracting logos from a graphical image etc. Most document analysis algorithms are developed by taking advantage of the underlying binarize image data. Degradations in document images result from poor quality of paper, the printing process, ink blot and fading, document aging, extraneous marks, noise from scanning, etc.

The goal of document binarization is to remove some of these artifacts and recover an image that is close to what one would obtain under ideal printing and imaging conditions. Binarization calculates the threshold value that categories stroke and background pixels. The use of two level information reduces the computational load compared to 256 levels of grey-scale or color image information. A binary image can be processed well and good than a grey scale image.

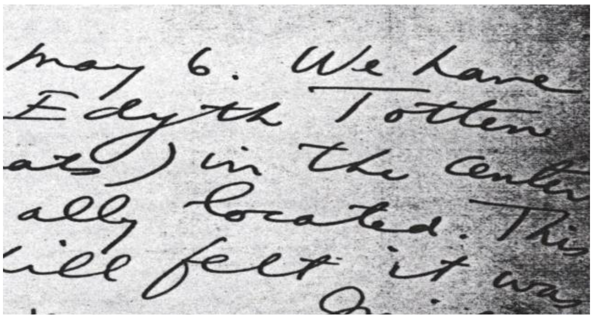
Historical documents are often degraded by different reasons as illustrated in Fig.1 (a) and (b).



platiť plné učiteľské platy.
má byť zriadená 3 členná rozhodčia
pp.dr.Oskar Reiss,ing.V.Holz a riad.

:Rozhodnuté bolo poštou podať m e m o
.a potom každá strana si menuje po
ať./Viď tu I 2/g./
rokov.a zákaz cestovania -

(a)



may 6. We have
Edyth Totten
at) in the cent
ally located. This
ill felt it wa

(b)

Fig.1 Degraded document image examples (a) and (b) are taken from DIBCO2011 series dataset

Thresholding is very important technique used in document Binarization. Thresholding has major two types, one is global thresholding and another is local thresholding. The global thresholding method computes an optimal threshold or single threshold for the whole image, so these methods required less computation and can work well in simple document images. But this technique is not suitable in case of complex documents. Local thresholding methods divide the image into sub-images blocks either statically or dynamically and then determine the threshold value for each block and convert it into binary image block depending on its local threshold value. Binarization process is already implemented with the help of various algorithms.Thesholding is very accurate and high speed approach to perform image segmentation. Thresholding is method to create binary image from gray scale image.

II. Related work

Many algorithms have implemented for the document image binarization from past decades. And still working on degraded document image is under process to generate more efficient, noiseless and clear document image. The binarization algorithms can broadly classify into two categories depending on choosing the technique of threshold values. Global thresholding algorithms in which a single threshold value is determined for the whole document image to binarize it.Examples of such algorithms are Otsu [2], Kittler and Illngworth [3], Kapur et al. [4], Papamarkos and Gatos [5]. Local thresholding algorithms in which a threshold for each pixel or a small region of the document images is determined to binarize the image. Examples of such algorithms are Bernsen [6], Niblack [7], and Sauvola and Pietikainen [8] belong to this category.There is one more category of binarization technique, which is called 'Hybrid Binarization techniques' where the techniques of both global and local thresholding approaches are combined. Many recent research works are going on 'Hybrid Binarization techniques' to combine the results of both global and local thresholding algorithms to produce better results by removing the drawbacks of both types of algorithms. So Hybrid Binarization algorithms are more popular to get good results. It is obvious to get better and good result by the proposed method than by canny operator, as the influence of the noise is restrained, meanwhile it is easy to get a segmentation result with higher precision.

III. Proposed system

In pre-processing of document image binarization is conversion of gray scale image from RGB image or color image. Gray scale image is essential for the elimination of noise, smoothing of background texture of degraded input document. Then on that gray scale image new image segmentation algorithm is applied to segment image into windows using threshold segmentation. In this method,each pixel in the image has its own threshold by calculating the statistical information of the grayscale values of its neighborhood pixels. According to threshold value gray scale image can binarize.

Some post-processing algorithm is applied on the binarize image. Foreground pixels that is separate from other foreground pixels are filtered out. Post-Processing can also connect break edges due to degradation. So we can generate more clear binarize image.

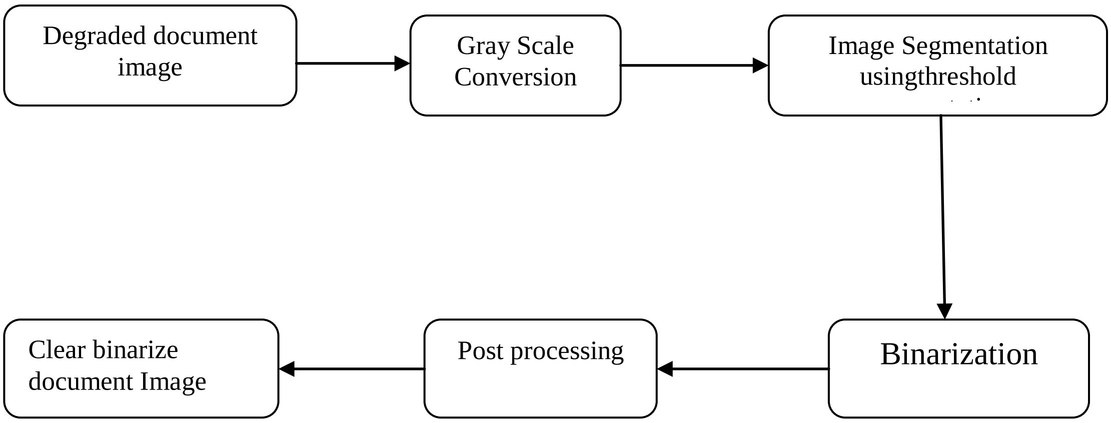


Fig.2 Degraded document image binarization using threshold segmentation

IV. Pseudo code

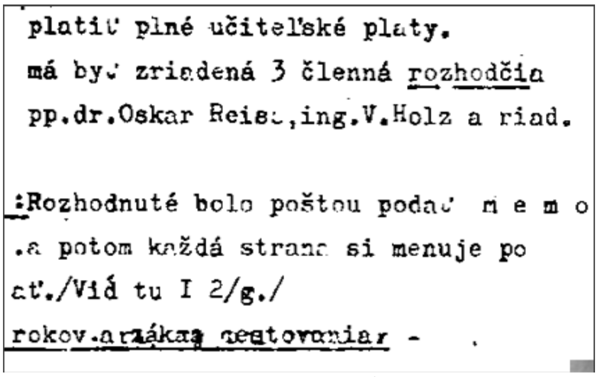
```
Image segmentation algorithm using threshold segmentation
Step 1:Input1.'G' is a gray scale image vector
            2. Set threshold value 'th'
            3. Set window size 'Ws'
                      4.'Bz' for binarize image vector
   Step 2:  For each row 1 to height-Ws
                  For column 1 to width-Ws
                      curr.pixel=G [row, column];
Step 3:  Check If (curr.pixel < avg-th)
                          Label Bz [row, column] =0;
                      else
                          Label Bz [row, column] =1;
                      end;
Step 4:  End
    Step 5:  Return Binarize image Bz.
```

V. Results

The proposed algorithm is implemented with C# in visual studio. Here we are giving degraded document image as an input to system. First it converts into gray scale image. Then proposed algorithm applied on gray scale image then binarization will perform and generates binarize image.

The figure 3 shows the binarize image after segmentation algorithm applied. Figure 3 shows the result after binarization, so this is nothing but partial result till we got. After this stage some post-processing technique can applied on the resulted image to produce clearer image. So that some break character stroke will try to connect through post-processing.

10.15680/ijircce.2015.0303099



platiť plné učiteľské platy.
má byť zriadená 3 členná rozhodčia
pp.dr.Oskar Reis.,ing.V.Holz a riad.

:Rozhodnuté bolo poštou podať memo
.a potom každá strana si menuje po
čť./Viď tu I 2/g./
rokov.a zákaz cestovania -

Fig.3.Binarize Image after Segmentation

VI. CONCLUSION

In this paper image segmentation using threshold segmentation for degraded document image binarization is proposed. The usage of less parameter in the techniques makes it simple and robust.This is tolerant to different types of degraded images. The Post-Processing technique is useful to generate more clear document image due to connecting unclear edges. The proposed algorithm will takes little time to get a precise result when small size structural operator is selected. It can generate good result than canny edge operator.

REFERENCES.

1. Bolan Su, Shijian Lu, and Chew Lim Tan,'Robust Document Image Binarization Technique for Degraded Document Images', IEEE Transaction on Image Processing, Vol. 22, No. 4, April 2013.
2. N. Ostu,A thresholding selection method from gray-level histogram', In: IEEE Trans. Systems Man Cyber net SMC-8,pp.62-66, 1978.
3. J. Kittler, J. Illingworth,Minimum error thresholding', In: Pattern Recognition, Vol. 19, No. 1, pp. 41-47. 1986.
4. J.N. Kapur, P.K. Sahoo, and A.K. Wong, 'A new method for gray-level picture thresholding', In: using the entropy of the histogram,Computer Vision Graphics and Image Processing, Vol. 29,pp.41-47, 1986.
5. N. Papamarkos, B. Gatos,'A new approach for multithreshold selection', Computer Vision Graphics and Image Processing,Vol. 56, No. 5, pp. 357-370,1994.
6. J. Bernsen,'Dynamic thresholding of grey-level images', Proceeding 8th ICPR,pp. 1251-1255, 1986.
7. W. Niblack, 'An Introduction to Digital image processing', In: Prentice Hall, pp. 115-116, 1986.
8. J. Sauvola and M. Pietikainen,Adaptive document image binarization',In: Pattern Recognition 33,pp.225–236, 2000