January 2012

A Novel Noise-removal Technique for Document Images

Samir Malakar

Dept. of MCA MCKV Institute of Engineering, Howrah, malakarsamir@gmail.com

Dheeraj Mohanta

Dept. of IT, Dept. of CSE Jadavpur University, Kolkata, India, dheeraj@gmail.com

Ram Sarkar

Dept. of IT, Dept. of CSE Jadavpur University, Kolkata, India, ramsarkar@gmail.com

Mita Nasipuri

Dept. of IT, Dept. of CSE Jadavpur University, Kolkata, India, mitanasipuri@gmail.com

Follow this and additional works at: <https://www.interscience.in/ijcct>

Recommended Citation

Malakar, Samir; Mohanta, Dheeraj; Sarkar, Ram; and Nasipuri, Mita (2012) "A Novel Noise-removal Technique for Document Images," *International Journal of Computer and Communication Technology*. Vol. 3 : Iss. 1 , Article 12.

DOI: 10.47893/IJCCT.2012.1116

Available at: <https://www.interscience.in/ijcct/vol3/iss1/12>

This Article is brought to you for free and open access by the Interscience Journals at Interscience Research Network. It has been accepted for inclusion in International Journal of Computer and Communication Technology by an authorized editor of Interscience Research Network. For more information, please contact sritampatnaik@gmail.com.

A Novel Noise-removal Technique for Document Images

Samir Malakar

Dept. of MCA

MCKV Institute of Engineering, Howrah,
malakarsamir@gmail.com

Dheeraj Mohanta, Ram Sarkar, Mita Nasipuri

Dept. of IT, Dept. of CSE

Jadavpur University,
Kolkata, India

Abstract--For developing a high quality Optical Character Recognition (OCR) system removal of noise from the document image is an utmost important step. To make this possible, filtering plays a significant role. Although mean and median filters, the two well-known statistical filtering techniques, are used commonly but sometimes these filters may fail to produce noise-free images or sometimes may introduce distortions on the characters in the form of gulfs or capes. In the work reported here, we have developed a new filtering technique, called Middle of Modal Class (MMC), for smoothing the input images. This filtering technique is applicable for both the noisy and noise free text document image at the same time. We have also compared our results with mean and median filters, and have achieved better result.

1. INTRODUCTION

Image processing, a fast growing field in the present days, consists of image acquisition, enhancement, segmentation and analysis. Out of these, image enhancement techniques, which include filtering, binarization and de-noising, are essential for preprocessing of acquired images for almost every application domain.

OCR, a sub field of image processing and pattern recognition, is a great challenge to the researchers from long time before [1]. There are several issues in developing an OCR system. One such issue is image enhancement i.e. extraction of noise free text from a textual document image. To achieve this filtering plays a remarkable role. Here, we have developed a new filtering technique, which is applicable for both the noise free and noisy textual document image at the same time.

2. PREVIOUS WORK

Lot of works [2-13] have been done on filtering techniques, which are applicable for digitized text pages or photographs, video signals, medical images etc. In the works [2-4], a diffusion filtering technique (both linear and non-linear) has been implemented by solving an initial boundary value problem for the two dimensional diffusion equations with a special non-linear source. In another work [5], the author had

introduced a fuzzy operator for de-noising noisy images and in [6], the authors had used a fuzzy operator based filtering technique for smoothing image. In the work [7], the authors had applied a fuzzy median filter, and using this operator the authors [8], had designed an enhancement technique for corrupted image. In the works [9-13], different variations of median filtering techniques were used. In the work [9], the authors had shown different types of properties of median filter. They had proved that repeated application of non-median n th ranked-order operations will reduce any signal to constant. They also had proved that any 1-D signal will be reduced to a root i.e., signal that is invariant to further filtering, after one-pass of recursive median filter. These properties were dedicated on signal processing. But in [10], the authors had discussed a new recursive median filter technique and had shown that their approach was guaranteed to converge to root with a finite number of iterations when it is applied on pseudorandom impulse noise. Their work was based on the functional optimization technique that states that median filtering is an optimization process in which two term cost function is minimized. In another work [11], the authors had designed a max/median filtering technique. In their technique they had first calculated median of each rows of 2-D window. Then they had taken the maximum of all these median values and had used in their max/median filter. In the work [12], the authors had provided a proof that any 2-D signal would converge to root after repeated passes of the recursive separable median filter and root structure differ those of their non-recursive counterpart. On this basis they had designed a separable median filtering technique. In another work [13], the concept of multistage filtering was introduced. In that work the window was divided into different subfield and then had filtered the window preserving a fixed property in a fixed direction.

3. MOTIVATION

All the works [2-13] have been applied on manually generated noisy images or signals. They even did not apply their techniques on noise free images. To have an automated system we have to apply a filtering technique on

both noise free and noisy document images, otherwise an extra overhead would be imposed for classification of noisy and noise-free document images. Also we need to classify different types of noises like Gaussian noise, Poisson noise, Speckle noise, Localvar noise, etc. These will obviously increase the computational complexity. So we severely required a technique, which is equally applicable for all types of images i.e., noisy or noise-free image at the same time. Although mean and median filter techniques are used commonly, sometimes these filters may fail to produce noise-free images or introduce some distortions on the characters in the form of gulfs or capes. In the light of the above-mentioned issues, we have introduced here a new filtering technique, which is equally applicable to both on noisy and noise-free textual document images simultaneously.

4. PRESENT WORK

In our present work we have introduced a new filtering technique, called MMC, for smoothing input images. A square window is considered which is slid over the whole document in raster order and for each position of the window, the gray level value of as the central pixel is replaced by a new value detected by the MMC technique. To get a better feel for the effect of noise filtering, the document image is finally binarized using a global threshold value.

We have also compared our results with two well-known statistical filtering techniques, namely equally weighted average or mean filter [14], and median filter [14] and have achieved better result. A flow diagram of the present work is shown in Figure. 1.

4.1. Window selection

For all the three filtering techniques mentioned above, we have considered a window of size $M \times M$ pixels, where „M“ is an odd positive integer number. Each square of the window are assigned with equal weight. We have applied our MMC filtering technique on each window and we have slid the window over the document image from left to right and from top to bottom and for each window position we have applied the MMC filtering techniques in order to achieve enhanced document image.

4.2. Filtering technique

In our technique we have found maximum occurrence of gray level values i.e., modal gray level value in a selected window. Now at the time of selecting the modal gray value for filtering it may so happen that all the pixels in the window may contain different gray level values as there are 256 different gray values. So to overcome this problem first we have searched the maximum (say, b) and minimum (say, a)

gray level values in a document image and have partitioned the interval $[a, b]$ equally, subject to the window size. Then we have applied pigeonhole principle [15]. In this principle it is stated that if m pigeons fly into n pigeonholes ($m > n$), then at least one pigeonhole will contain two or more pigeons. Since we have selected window size as $M \times M$ pixels, therefore we have partitioned the interval into $(M \times M) - 1$ equal partitions keeping pigeonhole principle in mind, each of length $(b-a) / (M \times M) - 1$. We have called each partition as modal class and length of partition as range of the class.

Figure 1. Basic flow diagram for the present work

Figure 1. Basic flow diagram for the present work

After that, we have put all the gray level values of the selected window into the proper class using the range information of the classes. Then we have searched for the class with maximum occurrence of gray level values i.e. the class with highest frequency and called it as modal class (MC). Next we have calculated the middle value (Midval) of the range of modal class i.e., if the range of MC is $[m, n]$ then $\text{Midval} = \lfloor m + n \rfloor / 2$. Finally, we have modified the central gray level value of the selected window with this Midval.

4.3. Binarization

Binarization is as important as that of filtering techniques in a document image processing. It helps to achieve clear document image and finally helps in further processing of image documents. There are different types of binarization techniques like, Otsu, Adaptive, Sauvola, Global threshold based, etc. Out of these binarization techniques we have implemented a global threshold-based binarization technique

and have shown the effect of all these three filtering techniques. The threshold value is estimated by calculating the mean of all modified gray level values of all the pixels in a document image.

5. EXPERIMENTAL RESULT AND DISCUSSION

For experimentation, we have classified the document images into 2 categories, such as noise free text document images, and noisy text document images. We have collected 50 text document images of each type for our experimental procedure. Some of the sample images are collected from DIBCO (Document Image Binarization Contest) 2009 [16] and rests are prepared in the CMTER laboratory, Dept. of CSE, Jadavpur University.

We have chosen the size of the window as 3x3 pixels, 5x5 pixels and 7x7 pixels. The results for some sample images are shown in Figure 2. (a-d). Visual observation of these sample results includes that our technique is giving better result for smaller window size i.e., window size of 3x3. We have used this window size to compare our techniques with the other two i.e., mean and median filter.

From the results shown in the Figure 3 (a-d), it's clear that our technique is delivering better result. We have compared our technique by visual inspection. We can't provide Mean Square Error (MSE) result as the original noise-free text document images are not available.

6. CONCLUSION

Image enhancement technique, which includes filtering, binarization etc. are essential for preprocessing of acquired images for almost every application in image processing domain. In the work reported here, a new filtering technique, MMC, is introduced for smoothing the input images. The filtering technique is applicable for both the noisy and noise free text document image at the same time. A comparison of our results with mean and median filters is done, and it is evident from the output images that the developed methodology outperforms better result. In the future development of the work, inclusion of more number and varieties of document pages, window size optimization and improved binarization techniques would be included.

7. ACKNOWLEDGEMENT

Authors are thankful to the "Center for Microprocessor Application for Training Education and Research", "Project on Storage Retrieval and Understanding of Video for Multimedia" of Computer Science & Engineering Department, Jadavpur University, for providing infrastructure facilities during progress of the work. The work reported here, has been partially funded by DST, Govt. of India,

PURSE (Promotion of University Research and Scientific Excellence) Programme.

REFERENCES

- [1] R.G. Casey, and E.Lecolinet, "A Survey of Method and Strategies in Character System", *IEEE TPAMI*, vol. 18, No. 7, July, 1996, pp. 690-706.
- [2] P. Perona, and J. Malik, "Scale Space and Edge Detection Using Anisotropic Diffusion", *IEEE TPAMI*, vol. 12, No. 7, 1990, pp. 629-639.
- [3] J. Weickert, "A Review of Non-linear Diffusion Filtering", *Lecture Notes on Computer Science*, vol. 1252, 1997, pp. 3-28.
- [4] G. V. Borisenko, and A. M. Denisov, "Nonlinear Source in Diffusion Filtering Methods for Image Processing", *Computational Mathematics and Mathematical Physics*, Vol. 47, No. 10, 2007, pp. 1631-1635.
- [5] F. Russo, and G. Ramponi, "Fuzzy operator for Sharpening of Noisy Images", *IEEE Electronics Letters*, vol. 28, No. 18, Aug., 1992, pp. 1715-1717.
- [6] F. Russo, and G. Ramponi, "Nonlinear Fuzzy Operators for Image Processing", *Signal Processing*, vol. 38, Aug., 1994, pp. 429-440.
- [7] A. Taguchi, H. Takashima, and Y. Murata, "Fuzzy Filters for Image Smoothing", in *Proc. SPIE Con., Nonlinear Image Processing V*, San Jose, CA, Feb. 7-9, 1994, pp. 332-339.
- [8] F.Russo, and G.Ramponi, "A Fuzzy Filter for Images Corrupted by Impulse Noise", *IEEE Signal Processing Letters*, Vol. 3, No. 6, June, 1996, pp. 168-170.
- [9] T. A. Nodes, and N. C. Gallagher, "Median filters: Some Modifications and Their Properties", *IEEE Transaction Acoustics, Speech and Signal Processing*, vol. ASSP-30, Oct., 1982, pp. 739-746.
- [10] G. Qiu, "An Improved Recursive Median Filtering Scheme for Image Processing", *IEEE Transaction on Image Processing*, vol. 5, no. 4, April 1996.
- [11] G. R. Arce, and M. P. McLoughlin, "Theoretical Analysis of Max/Median Filters", *IEEE Transaction on Acoustics, Speech and Signal Processing*, vol. ASSP-35, 1987, pp. 60-69.
- [12] G. R. Arce, and R. E. Foster, "Detail-Preserving Ranked-Order Based Filters for Image Processing", *IEEE Transaction on Acoustics, Speech and Signal Processing*, vol. ASSP-37, 1989, pp. 83-98.
- [13] M. P. McLoughlin, and G. R. Arce, "Deterministic Properties of The Recursive Separable Median Filter", *IEEE Transaction on Acoustics, Speech and Signal Processing*, vol. ASSP-35, Jan., 1987, pp. 98-106.
- [14] R. C. Gonzalez, and R. E. Woods, *Digital Image Processing*, first ed., Prentice-Hall, Indian, 1992.
- [15] http://en.wikipedia.org/wiki/Pigeonhole_principle , "Pigeonhole Principle".
- [16] <http://users.iit.demokritos.gr/~bgat/DIBCO2009/>, "DIBCO (Document Image Binarization Contest) 2009"

(a)

mädchen in das Schlafgemach der Dame führen, nahm
das an der Wand hangende Portrait ihres gerechten Lieb-
habers herab, und legte es auf ihr Coiffette, aus dem sie
Abends das Geld holen mußte, um ihren Diener zu be-

(c)

mädchen in das Schlafgemach der Dame führen, nahm
das an der Wand hangende Portrait ihres gerechten Lieb-
habers herab, und legte es auf ihr Coiffette, aus dem sie
Abends das Geld holen mußte, um ihren Diener zu be-

(b)

mädchen in das Schlafgemach der Dame führen, nahm
das an der Wand hangende Portrait ihres gerechten Lieb-
habers herab, und legte es auf ihr Coiffette, aus dem sie
Abends das Geld holen mußte, um ihren Diener zu be-

(d)

Figure 2. Illustration of noise filtering by MMC filter

(a) Original text documents

(c) Corresponding binarized documents after filtering with window size 5x5 pixels

(b) Corresponding binarized documents after filtering with window size 3x3 pixels

(d) Corresponding binarized documents after filtering with window size 7x7 pixels.

(a)

(b)

(c)

(d)

Figure 3. Illustration of various noise filtering techniques with window size 3x3 pixels

(a) Original text documents

(c) Corresponding binarized documents after filtering by Median filter

(b) Corresponding binarized documents after filtering by Mean filter

(d) Corresponding binarized documents after filtering by MMC filter