ISSN (Print) : 2320 – 3765
ISSN (Online): 2278 – 8875

International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering

*(An ISO 3297: 2007 Certified Organization)*

**Vol. 4, Issue 5, May 2015**

# Ancient Degraded Document Binarization Using Mean Shift Technique

P.S.Jonesherine, J.Paul Richardson Gnanaraj

M.E Student [Applied Electronics], Dept. of ECE, Kingston Engineering College, Vellore, Tamil Nadu, India

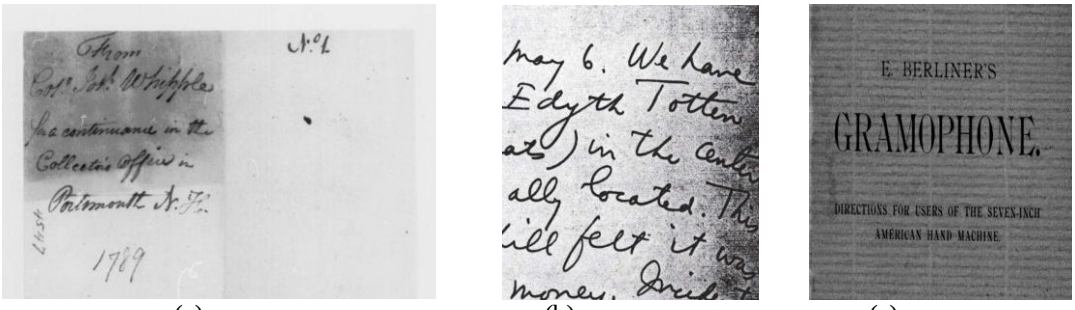
Assistant Professor, Dept. of ECE, Kingston Engineering College, Vellore, Tamil Nadu, India

**ABSTRACT:** In this paper a Mean shift algorithm is employed for ancient document images is proposed, as well as a post processing method that can improve any Binarization method. We introduce a local-global Mean Shift based colour image segmentation approach. It is a two-steps procedure carried out by updating and propagating cluster parameters using the mode seeking property of the global Mean Shift procedure. The first step consists in shifting each pixel in the image according to its R-Nearest Neighbour Colours (R-NCC) in the spatial domain. The second step process shifts only the previously extracted local modes according to the entire pixels of the image. Binarized model is made efficient by including mean shifting technique in the image. While in the post processing step, specialized adaptive Gaussian and median filters are considered. The result shows the output binarized image with the removal of global bleed through and few other degradation and proposed method is more efficient and provide better computed PSNR values comparing to the prior art.

**KEYWORDS:** Historical document Binarization, mean shift technique, document enhancement.

## I. INTRODUCTION

Libraries and archives around the world store an abundance of old and historically important documents and manuscripts. These documents accumulate a significant amount of human heritage over time. However, many environmental factors, improper handling, and the poor quality of the materials used in their creation cause them to suffer a high degree of degradation of various types. Today, there is a strong move toward digitization of these manuscripts to preserve their content for future generations. The huge amount of digital data produced requires a automatic processing, enhancement, and recognition. A key step in all document image processing workflows is binarization, but this is not a very sophisticated process, which is unfortunate, as its performance has a significant influence on the quality of OCR results.




(a) (b) (c)

Fig 1. Sample document images

Many research studies have been carried out to solve the problems that arise in the Binarization of old document images characterized by many types of degradation including faded ink, bleed-through, show-through, uneven illumination, variations in image contrast, and deterioration of the cellulose structure[1]. Fig. 1 shows some of the degraded document images used in this paper.

ISSN (Print) : 2320 – 3765
ISSN (Online): 2278 – 8875

International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering

*(An ISO 3297: 2007 Certified Organization)*

**Vol. 4, Issue 5, May 2015**

Mean shifting technique that uses the feature space to provide better Binarization and it provides better results for various types of degradation. The rest of the paper is organized as follows. In section II, we describe system analysis. In section III, the block diagram of the proposed Binarization model is presented, followed by a description of each step of the Binarization chain and a discussion of its impact. Section IV and V provides the mean shifting technique and comprehensive experimental results of our Binarization method respectively. Finally, section VI presents the conclusion about Binarization.

## II. SYSTEM ANALYSIS

A. *The global thresholding techniques:*

It computes an optimal threshold for the entire image these techniques need few computations and can work well in simple cases. But poor performance in complex backgrounds, such as non-uniform colour and poor illuminated backgrounds. These methods are usually not suitable for degraded document images, because they do not have a clear pattern that separates foreground text and background.

B. *The local Binarization techniques:*

It sets different thresholds for different target pixels depending on their Neighbourhood local information. Generally, these techniques are sensitive to background noises due to large variance in case of a poor illuminated document or bleed through degradation.

C. *Hybrid Binarization approach:*

This combines global and local thresholding. A first step consists in carrying out a global thresholding to classify a part of the background of the document image and keep only the part containing the foreground (graphics or text in our case). A second step aims to refine the image obtained by the previous step in order to obtain a sharper result by applying an adaptive thresholding Technique.

D. *Dynamic thresholding techniques:*

It is a iteration method that defines the threshold of a pixel with the grey-level values of its own and neighbouring pixels and the coordinate of the pixel. This Binarization method is commonly used for the bad quality images, especially the images with single peak histogram. Whereas the existing system may results in

- Less efficient for global bleed through removal
- Low efficiency of output image

## III. PROPOSED SYSTEM

*BINARIZATION MODEL*

The Final binarized output image is obtained by processing the input image in three steps: Preprocessing, Main Binarization, and Post Processing. Fig 2 gives the complete block diagram of binarization.The Binarization model is an extended version of the one proposed [2]

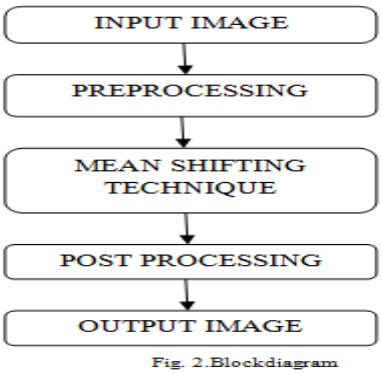



Fig. 2.Blockdiagram

ISSN (Print) : 2320 – 3765
ISSN (Online): 2278 – 8875

International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering

*(An ISO 3297: 2007 Certified Organization)*

**Vol. 4, Issue 5, May 2015**

A. *INPUT IMAGE:*

The natural scene image for text detection we give as input image.It may be any degraded document.

B. *PREPROCESSING:*

*Denoising:*

Denoising of images is typically done with the following process: The image is transformed into some domain where the noise component is more easily identified, a thresholding operation is then applied to remove the noise, and finally the transformation is inverted to reconstruct a noise-free image.fig 3 gives the flow of preprocessing. The wavelet transform has proved to be very successful in making signal and noise components of the signal distinct.

*Normalization:*

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. Then the otsu method is applied on the normalized image. The problem with this approach is that it misses weak strokes and sub-strokes, which means that we cannot rely on its output.

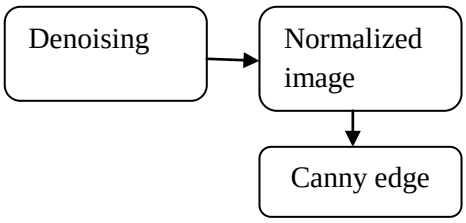



Fig. 3. Flow of Preprocessing

*Canny Edge Detection:*

To retrieve the better output the combination of binarized image with an edge mapobtained using the Canny operator [3].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. At the end of this step, the structure of foreground and text is determined. However, the image is still noisy, and the strokes and sub-strokes have not been accurately binarized. Also the binarization output is affected by some types of degradation.

C. *MEAN SHIFTING TECHNIQUE:*

The mean shift algorithm is a nonparametric clustering technique [4] which does not require prior knowledge of the number of clusters, and does not constrain the shape of the clusters. This paper focuses on reducing the computational cost in order to process large document images. We introduce thus a local-global Mean Shift based colour image segmentation approach. It is a two-steps procedure carried out by updating and propagating cluster parameters using the mode seeking property of the global Mean Shift procedure.

Step 1: It involves shifting each pixel in the image according to its *R-Nearest Neighbour Colours* (*R-NCC* ) in the spatial domain.
Step 2: This shifts only the previously extracted local modes according to the entire pixels of the image.

D. *POSTPROCESSING:*

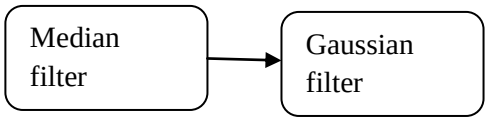



Fig. 4. Flow of Post Processing

*Median Filter:*

Median rejects salt and pepper noises in presence of edges.It preserves edges while removing noise. We use the median to remove the noise. After binarizing the input image using an adaptive median, bw Median, a simple

ISSN (Print) : 2320 – 3765
ISSN (Online): 2278 – 8875

International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering

*(An ISO 3297: 2007 Certified Organization)*

**Vol. 4, Issue 5, May 2015**

procedure is used to remove noises. By considering the binarized output image at this step, bw out, a connected component in bwout without any occlusion with bw Median is rejected from bw out foreground pixels.

*Gaussian Filter:*

Fig 4 gives the process flow in post process. Adaptive histogram equalization is applied on the binarized input image. Then adaptive Gaussian smoothing is applied on input image, bw Gauss. The bw Gauss map can reject an moderate amount of background pixels as well. However it introduces a little error on the interior pixels of large connected objects. To avoid this error bw Gauss is first filled by a morphological image filling method in the binary domain.

E. *OUTPUT IMAGE:*

Finally the output image has been obtained which is free from many types of degradation especially global bleed through and better binarization has been achieved.

## IV. MEAN SHIFT TECHNIQUE

### A. TYPES OF DEGRADATION:

*Foreground degradation:*

Text is nebulous, weak strokes/sub-strokes

*Background Degradation:*

*1) Bleed-Through :*

Bleed through degradation is an important and common interfering pattern [5] in the old and historical document images. Bleed through is categorized into two classes:

- local bleed-through
- global bleed-through

For removing local bleed-through, maximum moment of phase congruency covariance (MMPCC) is used.

*2) Unwanted lines/patterns:*

Interfering patterns and/or unwanted lines are common degradation types. While these types of degradation could be removed manually, It uses an adaptive median filter with a small scale to remove unwanted lines and patterns.

*3) Alien ink and faded ink:*

Usually faded ink is in the same colour of foreground text, but alien ink could be in any colours. To overcome these types of degradation, MMPCC and normalized denoised image are used, where values of denoised image are normalized between [0 255].

### B. DOCUMENT TYPE DETECTION:

To determine the type of input document we are dealing with. We propose to apply the enhancement processes that are after this step to the handwritten documents only and not to machine printed documents. We use the standard deviation of the orientation image that was produced during calculation of the phase congruency features. This image takes positive anticlockwise values between 0 and 180. A value of 90 corresponds to a horizontal edge, and a value of 0 indicates a vertical edge. By considering the foreground pixels of the output binary image obtained so far that the standard deviation value of the orientations for these pixels is low for handwritten document images and higher for machine-printed documents.

### C. MEAN SHIFT TECHNIQUE:

Image segmentation technique plays an important role in most image analysis systems. One of their major challenges is the autonomous definition of colour cluster number. Most of the works require an initial guess for the location or the number of the colours or clusters. They have often unreliable results since the employed techniques rely upon the correct choice of this number. If it is correctly selected, good clustering result can be achieved; otherwise, image segmentation cannot be performed appropriately.

ISSN (Print) : 2320 – 3765
ISSN (Online): 2278 – 8875

International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering

*(An ISO 3297: 2007 Certified Organization)*

**Vol. 4, Issue 5, May 2015**

In this paper, we propose an improved Mean Shift based two steps clustering algorithm. It takes into account a constrained combined feature space of the both colour and spatial information.

Our proposition has mainly three properties compared to the global Mean Shift clustering algorithm:

1) An adaptive strategy with the introduction of local constraints in each shifting process,
2) A combined feature space of both the colour and the spatial information,
3) A lower computational cost by reducing the complexity. Assuming all these properties, our approach can be used for fast pre-processing of real old document images. Experimental results show its desired ability for image restoration; mainly for ink bleed-through removal, specific document image degradation.

Mean shift mode finding process is illustrated in Figure 5. The mean shift clustering algorithm is a practical application of the mode finding procedure:

- starting on the data points, run mean shift procedure to find the stationary points of the density function,
- Prune these points by retaining only the local maxima.

The set of all locations that converge to the same mode defines the basin of attraction of that mode. The points which are in the same basin of attraction is associated with the same cluster.

1. *The global Mean Shift:*

The global Mean Shift clustering algorithm can be described as follows:

1. Choose the radius of the search window,
2. Initialize the location of the window $xj$ , $j = 1$,
3. Compute the Mean Shift vector $mh,G(xj)$,
4. Translate the search window by computing $xj+1 = xj + mh,G(xj)$, $j = j + 1$,
5. Step 3 and step 4 are repeated until reaching the stationary point which is the candidate cluster centre.

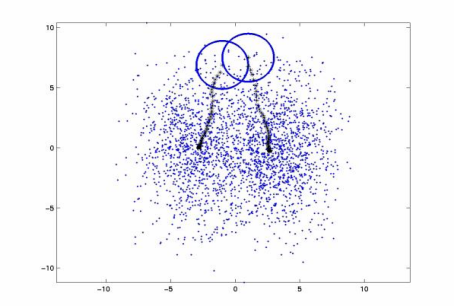


Fig.5. Mean Shift Mode Finding

2. *The local Mean Shift:*

The global Mean Shift algorithm, under its original form, defines a neighbourhood around the current point in the feature space related to the colour information. The neighbourhood refers to all the pixels contained in the sphere of a given arbitrary radius $\sigma R$ centred on the current pixel. It is extracted from a fixed size window and used for the *Parzen* window density estimation. Applying Mean Shift leads to find centroids of this set of data pixels. The proposed Mean Shift algorithm called the local Mean Shift algorithm is an improved version of the global Mean shift algorithm by reducing its complexity. Our main contribution consists in introducing a constrained combined feature space of the both colour and spatial information. Constraints are mainly introduced in the definition of a neighbourhood necessary for the estimation of the Mean Shift vector. Therefore, we introduce the concept of a new neighbourhood defined by the *R-Nearest Neighbour Colours*.

ISSN (Print) : 2320 – 3765
ISSN (Online): 2278 – 8875

International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering

*(An ISO 3297: 2007 Certified Organization)*

**Vol. 4, Issue 5, May 2015**

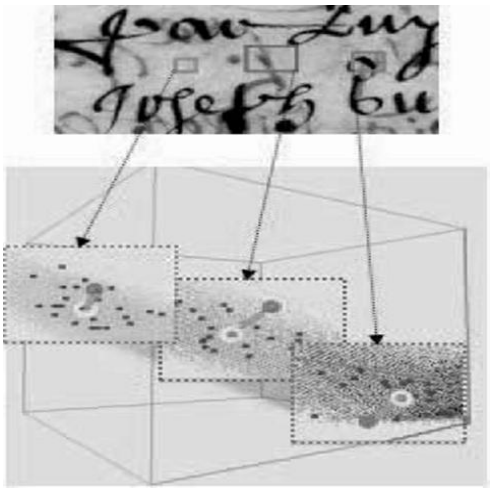


Fig.6. Scan of a manuscript and a zoom on a located window in the *L*u*v** cube after local Mean Shift application. Blue points are the *R* neighbours; red circle is a studied data image pixel; yellow circle is the extracted local mode.

It represents the set of the *R* nearest colours in the spatial domain extracted from an adaptive sliding window centred upon each studied data pixel in the image. *R* is an arbitrary predefined parameter. More precisely, we define the *R-NNC(X)* the *R* spatially nearest points from a given pixel *X* and having a colour distance related to *X* less than $\sigma R$. The studied neighbourhood of each pixel in the image, originally detected in a fixed window width, is modified in order to be defined from a gradually increasing window size. Starting from a 3x3 window size centred on a given data pixel *X*, we set for each neighbour *Y* within the window its colour distance from *X*. Then, we record all the neighbours having a colour distance less than an arbitrary fixed value $\sigma R$. If the number of the memorized data pixels is less than a fixed arbitrary value *R*, we increase the size of the window. We iterate the process of neighbours extraction and window increasing while the desired number of neighbours or the limit size of the window is not reached.The selection of the neighbours is as follow:

$$R - NNC(X) = \left\{ \begin{matrix} Y / d_{color}(X,Y) < \sigma_R \ is \ the \ spatially \\ nearest \ neighbor \ of \ X \end{matrix} \right\}$$

Intuitively, using here a progressive window size is of beneficial. This comes From the fact that computation of the mode is restricted inside a local window Centred on a given data pixel and more precisely restricted on the colourimetric ally and spatially nearest neighbours. By doing so, we guarantee an accurate convergence of the Mean Shift in few iterations. Figure 6 illustrates an example of the Mean Shift vector direction that points towards the direction of the most populated area. Furthermore, it is evident that the local mode closest to the value of the central pixel is a far better estimate of the true value than the average of all colour values.

3. *segmentation algorithm:*

The proposed segmentation algorithm follows the steps as below:

1. Run the local Mean Shift algorithm starting from each pixel *X* of the data set (converted to the feature space *L*u*v** ) and shifting over the *R-NNC(X)* neighbourhood. Once all the data pixels are treated, different local maxima of pixel densities are extracted.
2. Run the global Mean Shift algorithm starting from the extracted local modes and shifting over all pixels of the data image to reach the global maxima.
3. Assign to all the pixels within the image the closest previously extracted mode based on their colour distance from each mode. The number of significant clusters present in the feature space is automatically established by the number of significant detected modes.

ISSN (Print) : 2320 – 3765
ISSN (Online): 2278 – 8875

International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering

*(An ISO 3297: 2007 Certified Organization)*

**Vol. 4, Issue 5, May 2015**

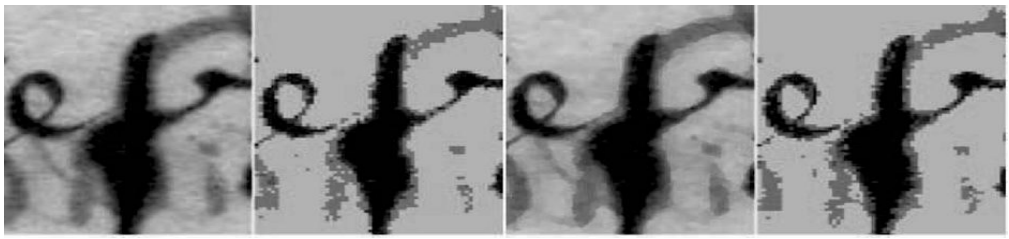


Fig.7. From left to right: an extract of a bleed-through degraded document, the segmented image with the global Mean Shift, the segmented image with the spatial Mean shift and the segmented image with the local-global Mean Shift

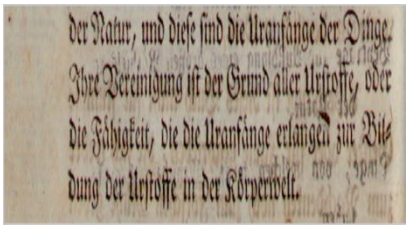


Fig.8 Input Degraded Document

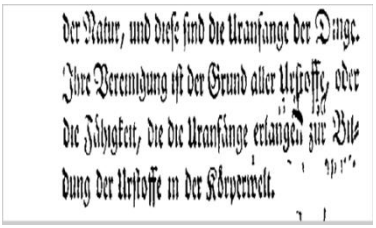


Fig.9.Binarized output image

## V. EXPERIMENTAL RESULTS

First we compare the subjective and objective performance of the proposed method with that of leading Binarization methods in the literature. Figure 7 explains how the bleed through has been extracted from the document. Then, we compared our proposed Binarization method with state-of-the-art algorithms and the top ranking algorithm in each competition and the obtained PSNR value is 37.66 which is good PSNR value comparing to other Binarization methods. An improved PSNR is obtained in this Binarization compared with many other methods. And the input degraded image and obtained binarized image is shown in figure 8and 9.

## VI. CONCLUSION

In this paper an image binarization method has been introduced that uses the information of the input image and employs the robust mean shifting technique and feature space extracted from that image are used to build a model for the Binarization of ancient manuscripts. This technique initially involves denoising followed by morphological operations is used to pre process the input image. Then mean shifting technique are used to perform the main binarization. For post-processing a median filter has been used to reject noise, unwanted lines, and interfering patterns. Because some binarization steps work with individual objects instead of pixels, a Gaussian filter was used.

## REFERENCES

[1] R. F. Moghaddam and M. Cheriet, **"A multi-scale framework for adaptive binarization of degraded document images,"** *Pattern Recognit.*, vol. 43,no. 6, pp. 2186–2198, 2010
[2] H. Z. Nafchi, R. F. Moghaddam, and M. Cheriet, **"Historical document binarization based on phase information of images,"** in *Proc. ACCV*,2012, pp. 1–12.
[3] J.Canny, **"A computational approach to edge detection,"** *IEEE Trans.Pattern Anal. Mach. Intell.*, vol. 8, no. 6, pp. 679–698, Nov. 1986
[4] D. Comaniciu and P. Meer**. Mean shift: A robust approach toward feature space analysis**. IEEE Trans. Pattern Anal. Machine Intell., 24:603–619, 2002
[5] I. B. Messaoud, H. Abed, H. Amiri, and V. Margner, **"New method for the selection of binarization parameters based on noise features of historical documents** .