An Adaptive Method for Physical Documents Digitization based on Global Energy Function Parameter

Ajay Kumar Pal
Research Scholar
Computer Science and Engineering Department
Samrat Ashok Technological Institute Vidisha (M.P.)

Yogendra Kumar Jain
Head of Department
Computer Science and Engineering Department
Samrat Ashok Technological Institute Vidisha (M.P.)

ABSTRACT
The first step of physical document analysis system is to digitalize the physical document. Recently number of researcher present numerous techniques that can vary in sensitivity, quality and some more control parameter. This paper presents a three tier framework for physical document digitization and describes an automatic technique for document digitization that can significantly increase the PSNR ratio.

General Terms
Image enhancement, Historical document preservation.

Keywords
Document digitization, Laplacian of image intensity, Canny edge, Gaussian filter, Markovian Random Field.

1. INTRODUCTION

Document digitization is an old but still a challenging and mind hunting task [1]. In the real world, appearance of printed documents may vary with quality of printing color and shadings, which degrade significantly quality of different binaries documents. Whereas the quality of different pixel of single binaries document may vary with light and view angle of different portion of physical document. Earlier one of the main goal of document digitalization is to differentiate the pixel of document image on the basis of their quality and take a proper treatment for them. But it is a very ambiguous process.

Digitations of physical documents are usually use a representation of 24-bit color, or may be 8 bits of grayscale [2]. In most of recent application, these representation techniques do not include all of the data available in original physical document, but retaining more than enough. Recently research leads to retain a single bit per pixel document. Credibility of digitations of physical document is relatively low, since information is loss, which is very much related to the primary information of the document. Many documents are produced using a monochrome ink for writing, and their meanings are incorporated exclusively for distribution of ink, a bit pattern representing the document explicitly.

Of course, the deduction of correct digitations of a document from colour or grayscale representation can be difficult. The physical deterioration of the document, image illumination conditions and limits of the unfavorable resolution can contribute to obscure the original pattern. Now these days many researchers proposed numerous algorithms for digitations of physical document towards this dispute. In fact standard document image binarization contest (DIBCO-11) held in-order to gather a deep research [3]. However, the results of these competitions prove that there is always room for enhancement in the quality of automatic binarization.

2. RELATED WORK

Otsu's method for document digitations is a parameter less global threshold method. In this method presence of separate distributions for background and text has been assume and calculates a threshold value in such a means that lead to minimize the difference between two distributions [4]. The limit for the distribution of two Otsu method was eliminated by Sezgin and Sankur [4], where the modes of degradation in the image histogram was removed by applying recursively otsu's method until only one mode remains in the picture. The overall limitation of the method is removed by Moghaddam and Cheriet [5] and an adaptation method is introduced, which uses the same concept as the Otsu method, but use local patches instead of global function for image.

Gatos et al. introducing a binarization method in which initially obtains a coarse binarization of the document image which is forwarded by rough background estimation [6]. Further local threshold values are calculated based on the estimated background. These threshold values are used to calculate the final binarization that is post-processed to remove noise and obtain the final output.

Gatos et al. presented another method [7]. This comprises four steps, first removing the background by polynomial fit of the lines, second detecting the contours of the race using Otsu's method on the gradient information then local threshold by averaging the detected pixels in a local neighborhood window edges, and finally post-processing of results.

Fabrizio and Marcotegui presented a method which is based on morphological mapping operator [8]. In morphological mapping, the analysis of those pixels are excluded whose erosion and dilation are too close, to avoid salt and pepper noise associated with it. Pixels are then classified as text, background and uncertain pixels. Then uncertain pixels are assigned to the text and background, depending on their boundary.

Although images of the documents may suffer from severe degradation and measure vastly, 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 [8, 10]. Su et al. proposed a frame which is using a binarization method to identify three classes; i.e. text, background and uncertain pixels, then uncertainty of pixels is reclassified using a classifier by using the classes of the text and background [11].

3. PROPOSED MODEL

Proposed framework for physical document digitations present three tier architecture where First, define the target binarization and labeling of pixels that minimizes global energy function based on a design of Markov random field model. Second, in the formulation of the length of data precision of the energy is based on the image intensity after being worked out using laplacian 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 all these points in more detail below with taking example of handwritten document i.e. HW3 from the dataset DIBCO-11.

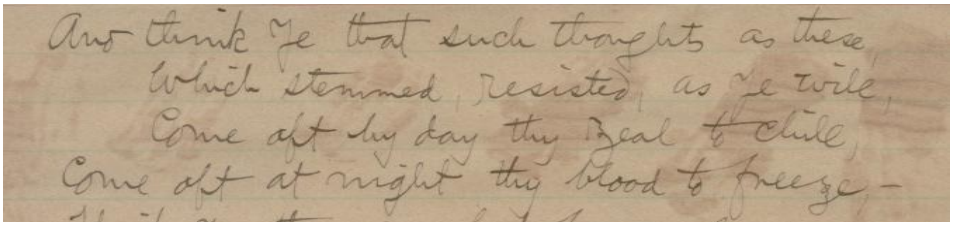

Fig 1: HW3 from DIBCO-11

3.1. Pixel Labelling

The global energy function works digitations D, label each pixel indexed by (i,j) is as ink or background, Dij ∈ {0, 1}. Energy comes as a additive, with terms that reflect the specific labeling fidelity compare to the current data, and corresponding terms representing the easeness or regularity of solution. In specifically, the energy includes a L^0ij or L^1ij cost understanding how the Di j label chosen for every pixel corresponds to its appearance, and irregular costs C^hij and C^vij for every pixel whose label respectively differ from its neighbor on both axis(horizontally and vertically).

$$\varepsilon I(D) = \sum_{i=0}^{m}\sum_{j=0}^{n}[L0ij(1-Dij) + L1ijDij] + \sum_{i=0}^{m-1}\sum_{j=0}^{n} Chij(Dij \neq Di+1, j) + \sum_{i=0}^{m}\sum_{j=0}^{n-1} C_{vij}(Dij \neq Di, j+1) \ldots\ldots (1)$$

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

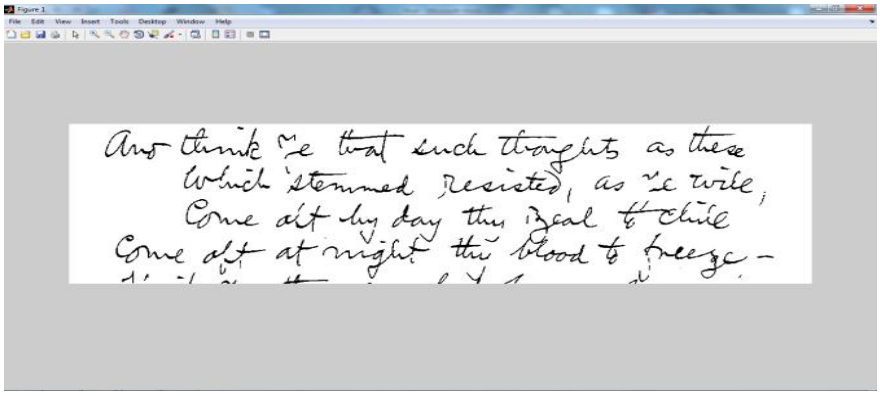

Fig 2: Using MRF (markovian random field) as Global Energy Function

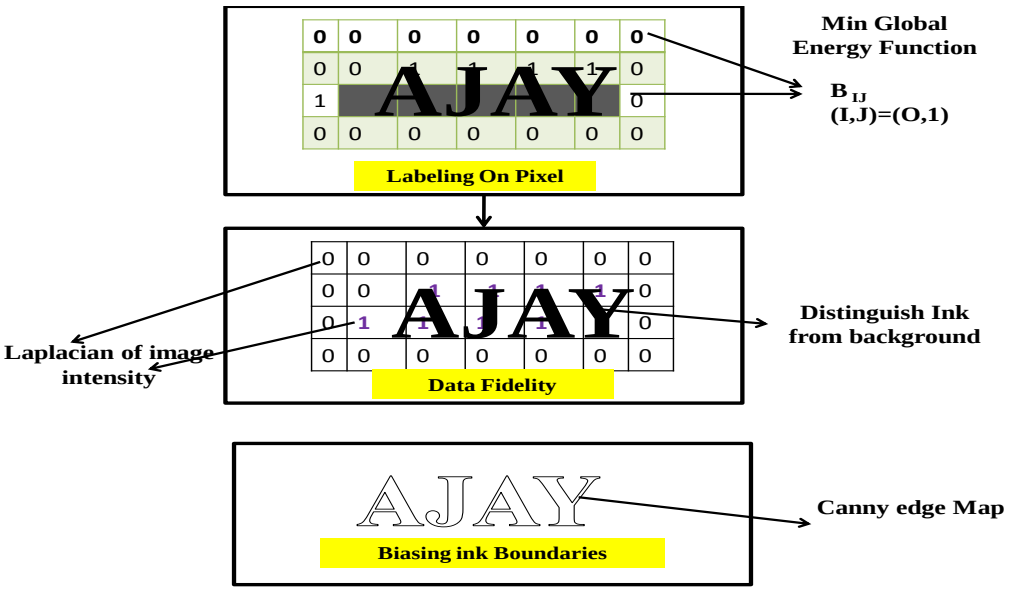

Fig 3: Proposed Frame Work for Document Image Digitations

3.2. Data fidelity

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 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 every component is calculated solely by what happens on its border. This makes some sense, but can create 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, it causes problems occasionally for background noise zones are eliminated 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 negative remark for all pixels (i, j) to the edge of the image, assuming that the ink is used 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 do not prevent the reorganization 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 this equation, φ has a large negative value.

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

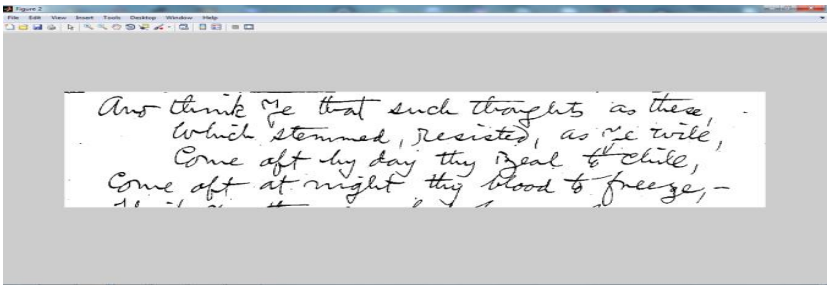

Fig 4: Distinguish Ink from background using Laplacian of Image Intensity.

3.3. Biasing Ink Boundaries

The mismatch penalties neighbor Chij and Cvij offer the possibility of using the third scheme mentioned above, using the method Canny .The algorithm describes the gap penalties to a scalar value c all over except between pair of 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 particular pixels. To address this problem, 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 converse choice will also make self-consistent, but will prevent the detection of broad strokes of single pixel.

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

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

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. Practically, this approach seems not much effective, because it only gives little motivation 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 very easy to select the place.

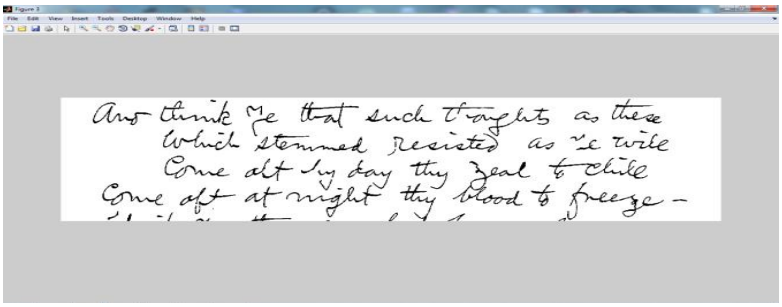

Fig 5: Final output after using Canny Edge detection

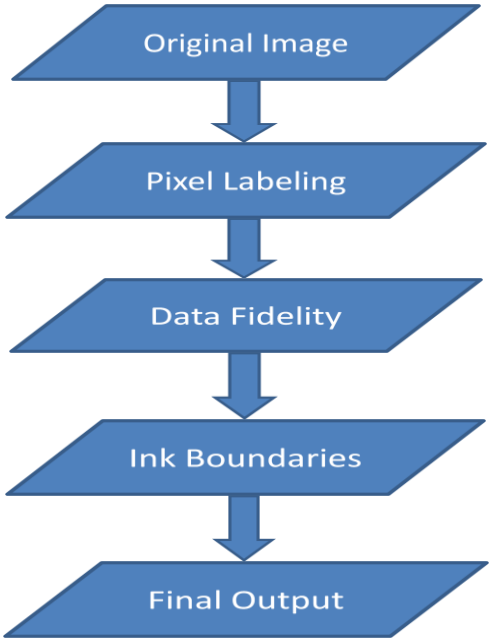

Fig 6: The proposed Three-tier architecture.

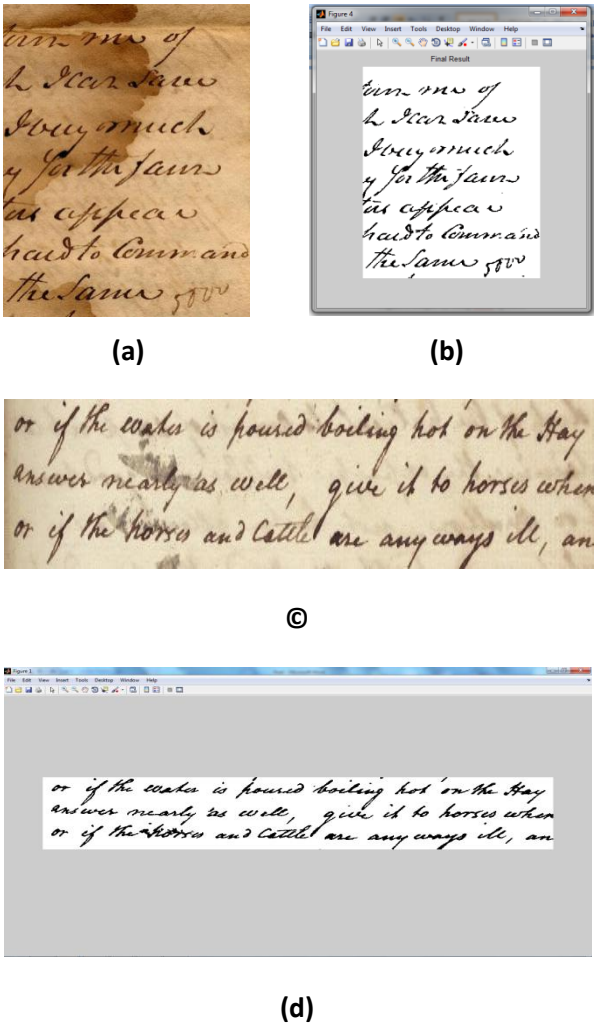

Fig 7: The performance of the proposed method for two historical document images from DIBCO-11. (a) and (c) are the input images HW4 and HW5 respectively. (b) and (d) are the outputs of the proposed method.

4. ENVIRONMENTAL SETUP & RESULT ANALYSIS

The proposed concept has been implemented in MATLAB. In this simulation, handwritten data set DIBCO11- has been used. This data set provides the input images as well as the output images which should be the final result. Results available of the DIBCO11 image dataset make possible a comparison between the existing algorithm and proposed method. The DIBCO11 contest used some additional primary criteria to describe the quality of the image. Of these, the peak signal-to-noise ratio (PSNR) correlates strongly with the quality of image. If G is the ground truth binary image and B is the binarization, then

$$PSNR = -10\log(\Delta(B,G))$$

In the proposed methodology we have applied three tier architecture to enhance the quality of the image. The step which includes biasing ink boundaries is carried out by using edge detection algorithm. To find out the best possible results, we have implemented different edge detection techniques to biasing ink boundaries and calculated respective PSNR values to obtain the best edge detection technique for the proposed method.

Sobel operator is the classical approach of finding the distorted edges by using 3x3 neighborhoods for the gradient calculation. The sobel operator is the magnitude of the gradient computed by:

$$M\sqrt{S_x^2 + S_y^2}$$

The next image is shown after applying both the masks on the HW4 during the ink boundaries. But there is a noticable difficulty it had on the edges of the certain characters, which are still distorted and unable to connect the true edges.

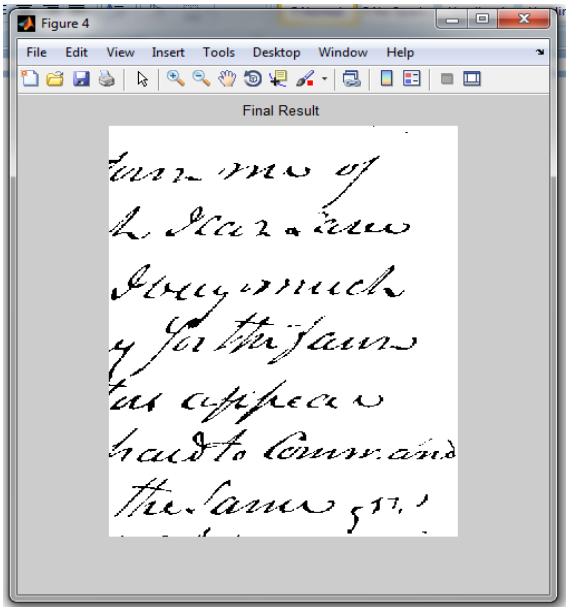

Fig 8: Output of HW4 image by including Sobel operator

Further in this work, the experiment is carried out by using canny edge detection algorithm. Canny algorithm computes the gradient magnitude using finite difference approximation for the partial derivatives. Along with it canny uses double thresholding to detect and link those edges. The Canny algorithm contains a number of adjustable parameters, which can affect the computation time and effectiveness of the algorithm.

- The Gaussian filter size: The smoothing filter which is used in the first stage affects directly the results of the canny edge detection algorithm. Smaller filters cause less blurring, allow small and sharp lines to be detected whereas a larger filter causes more blurring, smearing out the value of a given pixel over a larger area of the image. Larger blurring radii are more useful for detecting larger, smoother edges – for example, the edge of a rainbow.
- Thresholds: The two thresholds mechanism with hysteresis allows more flexibility than in a single-threshold approach. A threshold set too high can miss important information. On the other hand, a threshold set too low will falsely identify irrelevant information (such as noise) as important. It is difficult to give a single generic threshold that works well on all images. Therefore, the double threshold approach works well than the single threshold approach.

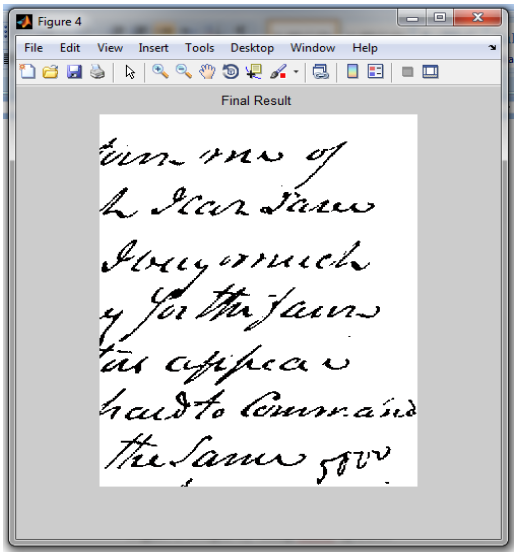

Fig 9: Output of HW4 by including Canny edge detection.

In similar way the figure 8 shows the output when the canny edge detection approach has been used. It can be seen that the output image above provides better detection and linkage of edges throughout the whole document as compared to the sobel operator.

The figure 9 shows the actual output which is provided by the data set DIBCO11-handwritten. To compare the results, we calculate the PSNR values of different handwritten documents from the dataset using both these edge detection algorithms separately. Shown below Table provid the PSNR values of canny and sobel operator. Column 2 represents the values of canny and column 3 represents the values of sobel.

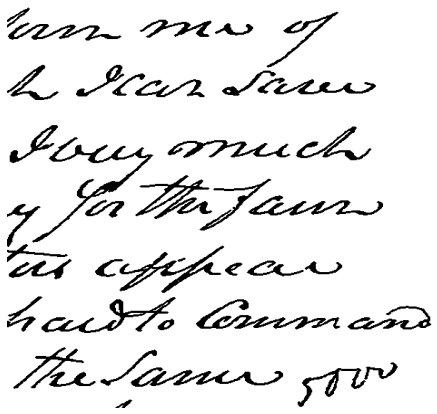

Fig 10: Output HW4_GT image provided by Result

The proposal is that higher the PSNR, the degraded image has been reconstructed in a better manner and able to match the original image and therefore better reconstructive algorithm. This would occur because we try to minimize the MSE (mean squared error) between images with respect the maximum signal value of the image. The mean squared error

for our practical purposes allows us to compare the "true" pixel values of our original image to our degraded image. Now, from the given below table, it can be easily said that the proposed algorithm works better with the canny edge detection algorithm.

Table 1: Comparison Results

Image Name	Canny	Sobel
HW1	23.4963	20.6584
HW2	60.4109	30.8057
HW3	46.5469	25.8506
HW4	42.5173	34.6834
HW5	47.497	39.0298

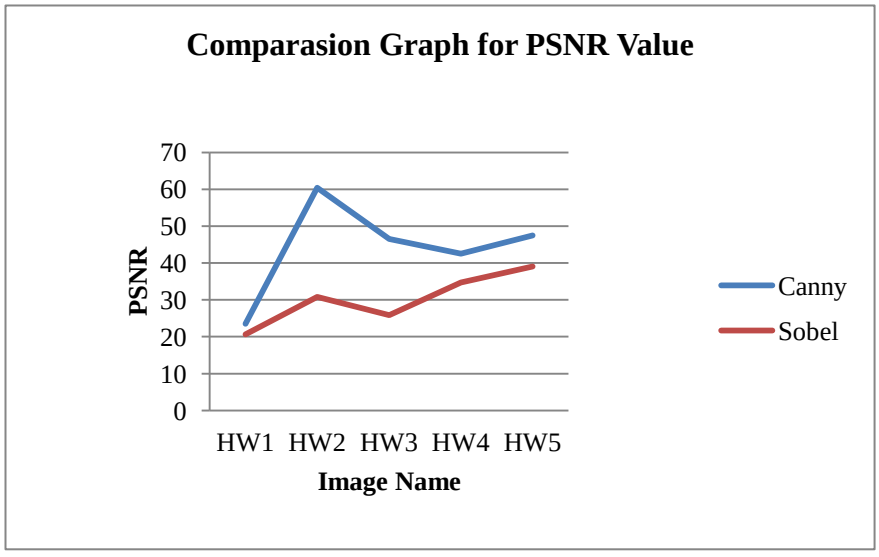

Figure 11: Comparison graph

5. CONCLUSION AND FUTURE WORK

In this dissertion, an adaptive binarization method inspired by Markov random field model is introduced. The first step was to label the pixels which comprise the global energy function. It then introduces a method which calculates the laplacian of the image intensity to distinguish between the ink from the background. Further it uses the canny edge detection algorithm to link the distorted edges and biasing the ink boundaries. We combined the global energy function with the edge detection to recover document images suffering from bleed through or intensity degradation. The number of single edges is reduced by almost half in the case of canny, thanks to its ability to recover weak and low intensity parts of the strokes on the edges. The method was evaluated on the DIBCO'11 dataset with promising results. This framework suggest to used luminous intensity of pixel and try to overcome the problem of appearance of a single document that can vary greatly depending on factors such as viewing angle, lighting. This method will help in preserving the historical documents with more accurate information for future use.

As a possible venue for future study, the other global methods such as entropy based methods can be used. Another future scope is to introducing the classifiers and combining them with edge detection methods to enhance the smoothness along the edges. Further improvements can be achieved, if try to implement an approach that can predict damage edge of character.

6. ACKNOWLEDGMENTS

I would like to thank **Dr. Yogendra Kumar Jain,** Head, Department of Computer Science and Engineering, who has contributed towards development of the template.

7. REFERENCE

[1] Reza Farrahi Moghaddamn , Mohamed Cheriet , 2012 "AdOtsu: An adaptive and parameterless generalization of Otsu's method for document image binarization", Elsevier transaction of Pattern Recognition, vol. 45, pp: 2419–2431.

[2] B. Gatos, K. Ntirogiannis, I. Pratikakis, 2009, "Document image binarization contest (DIBCO 2009)", International Conference on Document Analysis and Recognitation, pp: 1375–1382.

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

[4] M. Sezgin, B. Sankur, 2004, "Survey over image thresholding techniques and quantitative performance evaluation", Journal of Electronic Imaging, vol. 13, pp: 146–168.

[5] R. Farrahi Moghaddam, M. Cheriet, 2010, "A multi-scale framework for adaptive binarization of degraded document images", Elsevier transaction of Pattern Recognition, vol.43, pp: 2186–2198.

[6] B. Gatos, I. Pratikakis, S.J. Perantonis, 2006, "Adaptive degraded document image binarization", Elsevier transaction of Pattern Recognition, vol. 39, pp: 317–327.

[7] B. Gatos, K. Ntirogiannis, I. Pratikakis, 2010, "Document image binarization contest", International Journal on Document Analysis and Recognition, pp: 1–10.

[8] J. Fabrizio, B. Marcotegui, M. Cord, 2009, "Text segmentation in natural scenes using toggle-mapping", IEEE International Conference on Image Processing, pp: 2373–2376.

[9] B. Gatos, K. Ntirogiannis, I. Pratikakis, 2009, "Document image binarization contest", International Conference on Document Analysis and Recognition, pp: 1375–1382.

[10] R. Hedjam, R. Farrahi Moghaddam, M. Cheriet, 2011, "A spatially adaptive statistical method for the binarization of historical manuscripts and degraded document images", Elsevier transaction of Pattern Recognition, vol. 44, pp: 2184–2196.

[11] B. Su, S. Lu, C.L. Tan, 2010, "A self-training learning document binarization frame work", International Conference on Pattern Recognition, pg no: 3187–3190.