Hindi Character Segmentation in Document Images using Level set Methods and Non-linear Diffusion

Manjusha K, Sachin Kumar S, Jolly Rajendran, K. P. Soman
Centre for Excellence in Computational Engineering and Networking
Amrita Vishwa Vidyapeetham, Amrita School of Engineering, Coimbatore-641112

ABSTRACT

Hindi is the national language of India, spoken by more than 500 million people and is the second most popular spoken language in the world, after Chinese. Digital document imaging is gaining popularity for application to serve at libraries, government offices, banks etc. In this paper, we intend to provide a study on character binarization and segmentation of Hindi document images, which are the essential pre-processing steps in several applications like digitization of historically relevant books. In the case of historical documents, the document image may have stains, may not be readable, the background could be non-uniform and may be faded because of aging. In those cases the task of binarization and segmentation becomes challenging, and it affects the overall accuracy of the system. So these processes should be carried out accurately and efficiently. Here we experiment level set method in combination with diffusion techniques for improving the accuracy of segmentation in document process task.

Keywords

Level Set Method, Binarization, Segmentation, Convex Optimization.

1. INTRODUCTION

Digitization of documents helps in preserving books that are of great interest. If the original documents quality is good, then the digitized copy will retain it. Some of the copies might have got affected by different degradation over the time, through human manipulation, holes, spots etc. This opens an opportunity of research into Image Restoration techniques. These degradations can be considered as noises and so the problem finds the solution through noise removal. The restoration methods can really help in improving the accuracy of Optical Character Recognition (OCR). Conventionally, filters like average filter, median filter and gaussian filter are used for denoising. These filters reduce noise at the cost of smoothing the image and hence the edges will be softened. By using Partial Differential Equation (PDE) based methods, this issue can be solved. PDE based methods assume that the intensity of illumination on edges varies like geometric heat flow in which heat transforms from a warm environment to a cooler one until the temperature of the two environments reaches a balanced point [1]. These changes are in the form of Gaussian function [2]. PDE formulations are based on image derivatives. This approach considers images as piece wise continuous and differentiable functions [3]. The paper [4, 5] discusses two methods which are PDE based Active Contours. The basic idea of the Active Contours is to evolve a curve in a given image, and the evolution should stop when the curve meets an object or boundary. They are classified as parametric active contours and geometric active contours [6]. Geometric active contours are represented implicitly as level sets and this evolves according to Eulerian formulation. The Acitve Contour Model (ACM) [7-12] can be used to get binarized image.

Image binarization, conversion of a gray-scale document image into a binary image-is usually performed in the pre-processing stage of different document image analysis and applications. RGB or grey level images contain different intensity levels which make image processing difficult. Image binarization converts images of up to 256 gray levels to bi-level (black & white) images which reduces the amount of information. Simple binarization results in loss of information. Binarization should be computationally efficient and preserve the essential information. Different image binarization techniques [13-17] have been developed by many authors. Paper appearance, texture, stains, bleed-through effect, degradations caused by exposure to humidity or light complicates binarization of images.

The binarization techniques are classified into two- global and local. In global technique, same threshold value is used throughout the document. It doesn't take into account of the local information. So global binarization methods are not acceptable in many challenging cases. Therefore recent binarization methods use local thresholding method. They fall into two categories viz. Color intensity methods and edge based methods. Intensity based methods are again splitted up into threshold based and clustering based. In order to avoid the computational complexity of local methods, Grid based methods are used. Even though there are many binarization algorithms based on the above mentioned methods, here we make use of level set based approach which uses gradient for detecting edges. In the case of degraded images we make use of nonlinear diffusion along with level set based approach for getting better performance.

Segmentation results in the extraction of the individual objects in an image. PDE based models are found to be very effective upcoming methods for segmentation [7]. Active contour model is an effective model for segmentation [8-12]. The basic understanding of the active contour techniques through region based modelling can be understood from Mumford-Shah model [18]. There are other models such as Geodesic Active Contour (GAC) model/ Edge based model [7] in which sudden change in the image is taken as an edge/boundary. Gradient of image is used to identify the edges. In this paper we will discuss the effectiveness of two new/recent methods based on region based model which solves through convex optimization techniques. Region based method tries to find a separation of image objects with its background through an information calculated using image foreground and background. This will help the contour to gradually evolve on to the object.

This paper is organized as follows. Section 2 will give an overview about Hindi Language. In Section 3, we look into the level set theory, numerical form of Bresson's method (Level set method used for experimentation) and a brief overview of nonlinear diffusion. Section 4 deals with experimentation on Hindi document text images.

2. FEATURES OF HINDI LANGUAGE

India is a multi-lingual, multi-cultural, and multi-script country containing eighteen recognized languages and 15 scripts [19]. Hindi has its roots in Sanskrit language. Hindi is mostly written in Devanagiri script. Hindi is spoken using 52 sounds, 10 vowels, 2 modifiers and 40 consonants, nasalisation and a kind of aspiration. The vowels and consonants in Hindi language are shown below. The symbols shown below the alphabets are known as "matra" symbols. Matra symbols are used when consonants and vowels are to be written together [20, 21]. All the characters of Hindi language has a horizontal line at the upper part, known as 'Shirorekha' or headline. To form words using Hindi characters, these shirorekha's are connected to each other, thereby making it look like all the characters of the word are joined together by a common shirorekha.

3. BACKGROUND THEORY

Level set method is a numerical method which can follow the evolution of interfaces or it is a computational technique for tracking the propagating interface over time. In image segmentation, through level set method, a closed surface is evolved by expanding from a point, fitting the closed surface to the region it is released [22]. These interfaces can have sharp corners. This is a powerful tool in image and vision analysis for boundary detection. The key idea is to evolve a curve so that it will eventually stops along the object edges of the given image. The curve evolution is controlled by two sorts of energies: the internal energy defining the regularity of the curve and the external energy determined by the given image. The latter is often called the feature-driven energy.

To apply the level set method, a mesh or Cartesian grid is formed over the image. Level set method makes it very easy to follow shapes that change topology. Finer the mesh, more accurate is the level set. However, digitized image puts a limit on how fine a grid can be used. Because the image contains several thousands of pixels, and the grid must be as coarse as the individual pixels, and optimally even coarser. Once after the grid is formed, it applies a function $\phi(x,y,t)$, where (x,y) is a point in space at time 't'. This function is initialized to 0 at time t=0 and is called 'zero level set'. Thus in two dimensions, the level set method uses an auxiliary function ϕ to represent a curve C, and this function is called the level set function. The zero level set function can be represented as the curve C={(x,y)| $\phi(x,y)=0$}. ϕ has to be initialized at each point of the grid.

$$\phi(x,y,t) = \begin{cases} \text{dist}((x,y),C) & \text{if } (x,y) \in \text{ inside C} \\ 0 & \text{if } (x,y) \in \text{C} \\ -\text{dist}((x,y),C) & \text{if } (x,y) \in \text{ outside C} \end{cases}$$

The positive sign indicates that the point lie inside the curve and the negative sign indicates that the point is outside the closed curve and Figure.1 shows a pictorial representation. The curve evolves in time; the value of the function at each grid point also evolves. The level set method updates the value of the function at each point over small increments of time.
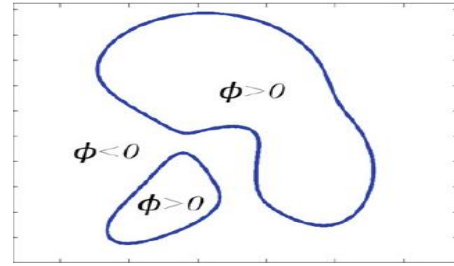

Figure 1: An arbitrary contour

In effect ϕ divides the image into 3 regions, inside of the level set with positive ϕ, boundary with $\phi=0$ and outside of the levelset with negative ϕ. Then, an iterative procedure is followed, which uses an edge stopping function to decide the rate at which, the curve evolves. The evolution of curve happens in a direction normal to itself and the evolution stops when the curve meets an object or boundary. Broadly there are two approaches- Edge based and Region based

3.1 Edge based model

In Edge based model the edge stopping function is defined as follows,

$$g(x,y) = \frac{1}{1 + |\nabla G_\sigma * f(x,y)|^2} \tag{1}$$

G_σ is a Gaussian filter is obtained from a normal distribution with standard deviation σ. When an edge is convolved with a Gaussian filter, diffusion occurs and the edge is slightly blurred. Near the edges denominator of $g(x,y)$ is very high. Thus $g(x,y)$ is a function which has small value near the edges and very high value when away from edges. So $g(x,y)$ acts as an edge stopping function [4].

3.2 Region Based Model

The Region based model is based on partitioning an image into object and background based on similarity of pixel intensities. It takes advantage of global regional information. Hence it is more robust than edge based model. The energy can be minimized using the following function

$$\min_{C,c_1,c_2} E(C,c_1,c_2) = \int_{inside(C)} (f-c_1)^2 dxdy +$$

$$\int_{outside(C)} (f-c_2)^2 dxdy + \lambda Length(C) \tag{2}$$

In equation 2, c_1 represents the average pixel value inside C and c_2 is the average pixel value outside C. λ is the regularization parameter.

Using implicit representation, it can be modeled as

$$\min_{C,c_1,c_2} E(C,c_1,c_2) = \int_{inside(C)} (f-c_1)^2 H(\phi) dxdy +$$

$$\int_{outside(C)} (f-c_2)^2 (1-H(\phi)) dxdy + \lambda \int_\Omega |\nabla H(\phi)| dxdy \tag{3}$$

c_1 and c_2 are calculated as

$$c_1 = \frac{\int\limits_\Omega f(x,y)H(\phi)dxdy}{\int\limits_\Omega H(\phi)dxdy}$$

$$c_2 = \frac{\int\limits_\Omega f(x,y)(1-H(\phi))dxdy}{\int\limits_\Omega (1-H(\phi))dxdy} \qquad (4)$$

Calculus of variation is used to find the optimal ϕ function by using the relation

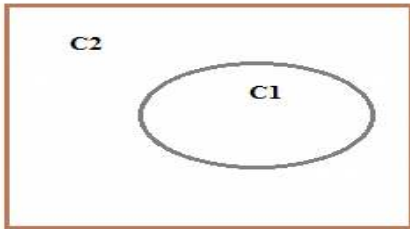$$\frac{\partial \phi}{\partial t} = -\frac{\partial E}{\partial \phi} \qquad (5)$$

Figure 2: Geometric representation of region based model

3.3 Fast Global Minimization Algorithm for Active Contour Models by Xavier Bresson[5]

It is one of the efficient minimization method for the active contour minimization problem. It is memory efficient and have better performance compared to other methods like graph cut technique. The energy minimization problem of active contours can also be written as

$$\min_c \{F_{AC}(C) = \int_C g_b(C,s)ds + \lambda \int_{C_{in}} g_r^{in}(C_{in},x)dx + \lambda \int_{C_{out}} g_r^{out}(C_{out},x)dx\}$$

$$(6)$$

This is equivalent to the energy minimization of region based model we discussed above. C represents the closed curve, C_{in} is the inside region, C_{out} is the outside region. g_b is the boundary function, g_r^{in} and g_r^{out} are inside and outside region function. The problem can be formulated as a convex optimization problem and can be minimized using Bregman's approach.

3.4 Non-linear Diffusion

Diffusion is a physical process that equilibrates concentration differences without creating or destroying mass.

The diffusion equation is

$$\partial_t u = \text{div}(D \cdot \nabla u) \qquad (7)$$

where ∇u is the concentration gradient and D is the diffusion tensor matrix. If we replace the diffusion tensor D by a single scalar value for the whole image domain, the process is equivalent to uniform smoothing. If D itself depends on the evolving image u, the resulting equation describes a nonlinear diffusion filter. Nonlinear diffusion has deservedly attracted much attention in the field of image processing for its ability to reduce noise while preserving (or even enhancing) important features of the image, such as edges or discontinuities; this can be opposed to linear diffusion which not only removes noise but also blurs and dislocates edges [25].

Nonlinear diffusion was first suggested in the pioneering work of Perona and Malik [27]. Motivated mainly by edge detection, they criticized the linear scale space for which the true location of a boundary at a fine scale is not directly available in the coarse scale image, edges are blurred and dislocated, and their tracking to finer scales is complicated and computationally expensive. In edge-enhancing anisotropic diffusion, the diffusivity function allows more smoothing parallel to image edges and less smoothing perpendicular to these edges. The algorithm is described in [24]. Coherence enhancing anisotrpic diffusion is an extension of edge-enhancing anisotropic diffusion that is specifically tailored to enhance line-like image structures by integrating orientation information [26].

4. EXPERIMENTS AND RESULTS

In this section we will show how the active contour model based methods do the binarization and segmentation process. The experiments are done in MATLAB. The images used for testing are collected from various Hindi books and newspapers and internet.

At first we can see how the contour is evolved on the Hindi text image. Figure 3 represents the input test image. Figure 4 represents the step by step contour evolution on that image. It starts from some arbitrary point and on each iteration the level set moves more close to the major edges in the image and finally it stops when it exactly reaches the outer boundary of Hindi words. Figure4 (f) represents the binarized form of finally evolved contour. Here we can see that binarized image is constructed by making the pixel value on the contour boundary as black and all other pixels white. Finally level set method gives a denoised (by considering only the relevant character boundaries), binarized segmented image.

Figure 5 is the third input image with text paragraph having three lines of text. It is given as input to the Bresson's level set method and the obtained result is as shown in Figure 6. Then for line segmentation, the information obtained from evolved contour can be used. The segmented lines are shown in Figure 7.

As Hindi language has 'Sirorekha' in all the words, to identify each character in the given word 'Sirorekha' has to be removed. By using histogram profiling character boundaries can be identified by getting the columns having lower number

होना चाहिए

Figure 3:Image1

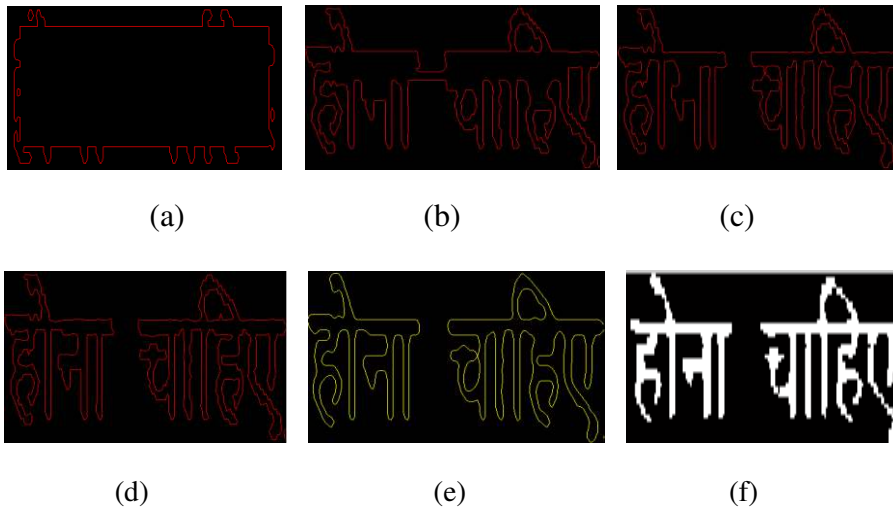(a)   (b)   (c)

(d)   (e)   (f)

Figure 4: Contour evolution using level set method on image1.

संघ के लिए होना चाहिए, संघ के एक सदस्य के लिए नहीं। उन्होंने बहुत
अनुनय-विनय की, परन्तु उन्होंने (बुद्ध ने) उसे स्वीकार करने से इनकार कर
दिया, वह बिलकुल नहीं माने।

Figure 5: Image of sentences

संघ के लिए होना चाहिए, संघ के एक सदस्य के लिए नहीं। उन्होंने बहुत
अनुनय-विनय की, परन्तु उन्होंने (बुद्ध ने) उसे स्वीकार करने से इनकार कर
दिया, वह बिलकुल नहीं माने।

(a)

संघ के लिए होना चाहिए, संघ के एक सदस्य के लिए नहीं। उन्होंने बहुत
अनुनय-विनय की, परन्तु उन्होंने (बुद्ध ने) उसे स्वीकार करने से इनकार कर
दिया, वह बिलकुल नहीं माने।

(b)

Figure 6: First image gives the contour form and the second image shows the binarized form

संघ के लिए होना चाहिए, संघ के एक सदस्य के लिए नहीं। उन्होंने बहुत

(a)

अनुनय-विनय की, परन्तु उन्होंने (बुद्ध ने) उसे स्वीकार करने से इनकार कर

(b)

दिया, वह बिलकुल नहीं माने।

(c)

Figure.7 Line segmented output of second figure in figure 6

अनुनय विनय की, परन्तु उन्होंने (बुद्ध ने) उसे स्वीकार करने से इनकार कर

Figure 8: On figure 7 (b) histogram is taken and identified character boundaries

अ नु न य वि न य की , प र न्तु उ न्हों ने (बु द्ध ने)

उ से स्वी का र क र ने से इ न का र कर

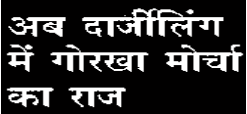**Figure 9: Character segmented output of figure 8**
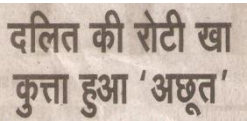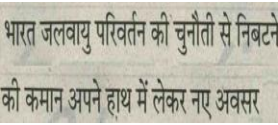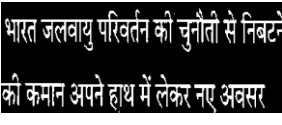
Input Image	Binarized Image	Number of wrongly segmented words	Time taken for word segmentation	Accuracy
हर प्रकार की हम्द व सना (प्रशंसा और गुणगान) अल्लाह के लिए योग्य है, हम उसी की प्रशंसा करते हैं, उसी से मदद मांगते और उसी से हमा याचना करते हैं, तथा हम अपने नफ्स की बुराई और अपने बुरे कामों से अल्लाह की पनाह में आते हैं, जिसे अल्लाह तआला हिदायत दे दे उसे कोई पथभ्रष्ट (गुमराह) करने वाला नहीं, और जिसे गुमराह कर दे उसे कोई हिदायत देने वाला नहीं। हम्द व सना के बाद :	हर प्रकार की हम्द व सना (प्रशंसा और गुणगान) अल्लाह के लिए योग्य है, हम उसी की प्रशंसा करते हैं, उसी से मदद मांगते और उसी से हमा याचना करते हैं, तथा हम अपने नफ्स की बुराई और अपने बुरे कामों से अल्लाह की पनाह में आते हैं, जिसे अल्लाह तआला हिदायत दे दे उसे कोई पथभ्रष्ट (गुमराह) करने वाला नहीं, और जिसे गुमराह कर दे उसे कोई हिदायत देने वाला नहीं। हम्द व सना के बाद :	2	12.69 sec	97.72%
दलित की रोटी खा कुत्ता हुआ 'अछूत'	दलित की रोटी खा कुत्ता हुआ 'अछूत'	0	3.22 sec	100%
अब दार्जीलिंग में गोरखा मोर्चा का राज	अब दार्जीलिंग में गोरखा मोर्चा का राज	0	1.42 sec	100%
भारत जलवायु परिवर्तन की चुनौती से निबटने की कमान अपने हाथ में लेकर नए अवसर	भारत जलवायु परिवर्तन की चुनौती से निबटने की कमान अपने हाथ में लेकर नए अवसर	0	1.77 sec	100%

Table.1 Experiments and it's Results

of non-zero pixels. Then by setting those columns to value zero, the obtained image is shown Figure 8. Figure 9 shows the segmented characters obtained when Bressons method is applied on Figure 8. The Bressons method is computationally efficient as it takes less time for the segmentation process. In Bressons method no need for applying binarization explicitly,

it automatically happens during the segmentation process itself.

Level set based Bresson's method is applied to several different images. The results are shown in Table.1. Character segmentation accuracy is calculated by finding the total number of characters in the input image and total number of correctly segmented characters. In Table 1, The results shown are obtained without applying any pre-processing operations on the image. When images have noise like Salt and Pepper, the Bresson's algorithm gives lower performance as the noise itself is detected as characters and it draws contour on each of these. Bresson's method gives low performance on those type of noises which have almost the same intensity value as that

of the objects inside the image. Figure 10 shows the result of Bresson's method in one of such images. The result shows poor performance. But when edge enhancement diffusion on to the input image, the noise is removed and the Bressons method gives good result. The result after applying the diffusion is shown in Figure 11. So the combination of nonlinear diffusion methods with the Bresson's level set based method improves the overall performance of segmentation in the case of highly degraded or highly noised images. The binarised output of the image obtained without applying diffusion and with applying diffusion is shown in the Appendix.
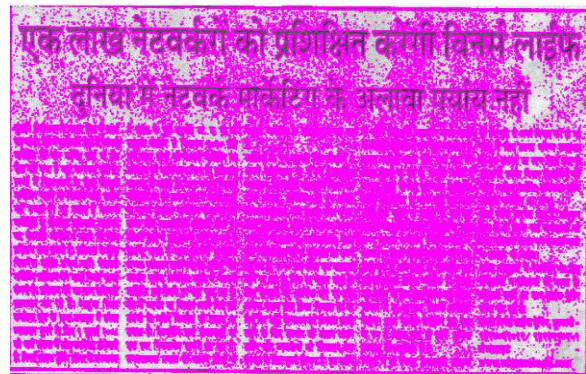

Figure 10: Image on left is the input image and on the right is contour formed image after applying bressons algorithm

Figure 11: Image on left is the input image and on the right is contour formed image by bressons algorithm, after applying edge enhancement diffusion

To demonstrate the good performance in segmentation of Bressons method along with nonlinear diffusion, an input image is taken and different levels of noise is added into

it.Those noisy images are given for the experiment. The results obtained are shown in Figure 12.
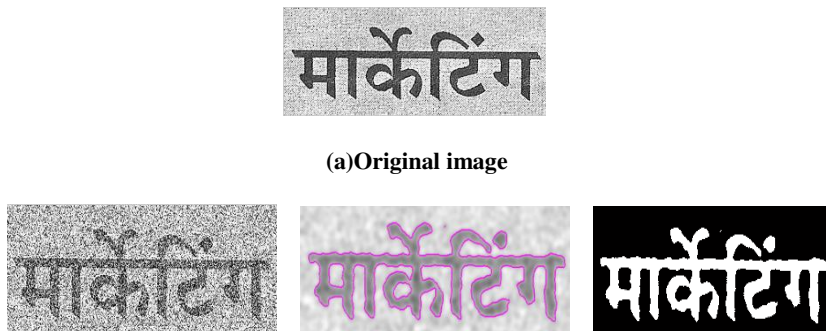
(a)Original image

(a) Gaussian noise is added to (a), is shown in first figure. The second figure represents the segmentation result obtained after applying diffusion and the third figure represents the binarized form of second figure

(b) Salt and pepper noise is added to (a), is shown in first figure. The second figure represents the segmentation result obtained after applying diffusion and the third figure represents the binarized form of second figure

(c) Speckle noise is added to (a), is shown in first figure. The second figure represents the segmentation result obtained after applying diffusion and the third figure represents the binarized form of second figure

Figure 12: Results of Diffusion with Bresson's Segmentation

5. CONCLUSION

In this paper, we studied how level set based Bresson's method, is used for image binarization and segmentation of Hindi character images. The nonlinear diffusion method helps in improving the segmentation performance when noises like salt and pepper, gaussian and speckle noise is added to the input image. The results obtained from experiment (done in MATLAB), shows that the Bresson's method takes less time for computation.

6. REFERENCES

[1] P.Perona and J. Malik, "*Scale-Space and Edge Detection using Anisotropic Diffusion*", IEEE Trans. Pattern analysis and Machine Intelligence, vol. 12, no. 7, pp. 629639, July 1990

[2] F. Drira, F. LeBourgeosis, H. Emptoz, "*A new PDE-based approach for singularitypreserving regularization: application to degraded characters restoration*", International Journal on Document Analysis and Recognition (IJDAR)(2011): 1-30, May 2011

[3] C. Xu, A. YezziJr, J.L. Prince, "*On the relationship between parametric and geometric active contours*", Asilmor Conf. Signals, Systems, and Computers, pp. 483-489, 2000

[4] Kaihua Zhang, Lei Zhang, Huihui Song, Wengang Zhou, "*Active contour with selective local or global segmentation: A new formulation and level set method*", Elsever, 2009

[5] Xavier Bresson, "*A Short Guide on a Fast Global Minimization Algorithm for Active Contour Models*", April 22, 2009.

[6] S.K. Weeratunga, C. Kamath, "*An Investigation of Implicit active contours for scientific image segmentation*", in: Visual Communications and Image Processing Conference, 2003

[7] M. Kass, A. Witkin, D. Terzopoulos, "*Snakes: active contour models*", International Journal of Computer Vision 1 (1988) 321–331.

[8] N. Xu, N. Ahuia, R. Bansal, "*Object segmentation using graph cuts based active contours*", Computer Vision and Image Understanding 107 (2007) 210–224

[9] V. Caselles, R. Kimmel, G. Sapiro, *Geodesic active contours*, in: Processing of IEEE International Conference on Computer Vision'95, Boston, MA, 1995, pp.694–699

[10] M. Ben Salah, A. Mitiche and I. Ben Ayed, "*Effective Level Set Image Segmentation with a Kernel Induced Data Term*", IEEE Transactions on Image processing, vol. 19, no 1, pp. 220–232, 2010.

[11] T. Chan, L. Vese, "*Active contours without edges*", IEEE Transaction on Image Processing 10 (2) (2001) 266–277

[12] G.P. Zhu, Sh.Q. Zhang, Q.SH. Zeng, Ch.H. Wang, "*Boundary-based image Segmentation*"

[13] J.Ohya, A. Shio, and S. Akamatsu, "*Recognizing characters in scene images*", IEEE Trans. Pattern Anal.Mach.Intell., 16(2), 1994, pp.214–220

[14] Y. Zhong, K. Karu, and A.K. Jain, "*Locating text in complex color images*", Pattern Recognition, 28(10) , 1995, pp.1523–1535

[15] O. D. Trier and T. Taxt, "*Evaluation of binarization methods for document images*", IEEE Trans. Pattern Anal. Machine Intell., vol. 17, Mar. 1995, pp. 312-315.

[16] .T. Abak, U. Baris, and B. Sankur, "*The Performance Evaluation of Thresholding Algorithms for Optical CharacterRecognition*", ICDAR 97, Ulm, Germany, 1997, pp. 697-700

[17] Mehmet Sezgin, "*Survey over image thresholding techniques and quantitative performance evaluation*", Journal of electronic imaging, 13, 146,2004, doi:10.1117/1.1631315

[18] R. Malladi, R. Kimmel, D. Adalsteinsson, G. Sapiro, V. Caselles, and J. A. Sethian. "*A geometric approach to segmentation and analysis of 3d medical images*", In MMBIA '96: Proceedings of the 1996Workshop on Mathematical Methods in Biomedical Image

Analysis(MMBIA '96), page 244, Washington, DC, USA, 1996.IEEE Computer Society

[19] C. V. Jawahar, M. N. S. S. K. Pavan Kumar, S. S. Ravi Kiran, *"A Bilingual OCR for Hindi-Telugu Documents and its Applications"*, ICDAR, vol. 1, pp.408, Seventh International Conference on Document Analysis and Recognition (ICDAR'03) - Volume 1, 2003

[20] U.Pal, B.B.Chaudhuri, *"Indian Script Character Recognition: A survey"*, Patter Recognition, vol. 37, pp. 1887-1889, 2004

[21] E. Nadernejad, H.Koohi, and H. Hassanpour, *"PDEs-Based Method for Image Enhancement,"* Applied Mathematical Sciences, Vol. 2, No. 20, pp. 981 – 993, 2008

[22] H. Philips, The Level Set Method, http://web.mit.edu/aram/www/work/thesis.pdf

[23] K. Kang, C. Weinberger, W. Cai, *"A Short Essay on Variational Calculu"s*, Dept of Mechanical Stanford University, May 2006

[24] Hyunwoo Kim, Jeong-Hun Jang, Ki-Sang Hong, *"Edge-Enhancing Super-Resolution Using Anisotropic Diffusion"*, ©2001 IEEE

[25] PavelMr´azek, *"Nonlinear Diffusion for ImageFiltering and Monotonicity Enhancement"*

[26] Joachim Weickert,, *"Coherence-Enhancing Diffusion Filtering"*, International Journal of Computer Vision 1999 Kluwer Academic Publishers.

[27] Perona, P. and Malik, J. *"Scale space and edge detection using anisotropic diffusion"*, IEEE Trans.Pattern Anal. Mach. Intell.,Vol. 12, pp. 629–639

7. APPENDIX

Figure 13: Binarized Output without applying non-linear diffusion

Figure 14: Binarized Output after applying non-linear diffusion