Level Set Methodology for Tamil Document Image Binarization and Segmentation

S. Karthik, Hemanth.V.K, V. Balaji, K. P. Soman,
Centre for Excellence in Computational Engineering and Networking, Amrita Vishwa Vidyapeetham, Coimbatore, Tamilnadu, India, 641112.

ABSTRACT
The most challenging task in OCR is getting the characters segmented properly. The accuracy of segmentation depends on the quality of the binarization technique applied. Binarization is the process of setting all intensity values greater than some threshold value to ”on”. It converts the document image into binary image as extracting text and eliminating the background. This process also removes the noise. The output of this process is used as input to image segmentation process. Conventionally separate methods are used for binarizarion and segmentation. In this paper we investigate the use of recently introduced convex optimization methods, selective local/global segmentation (SLGS) algorithm [16] and fast global minimization (FGM) algorithm [15] for simultaneous binarization and segmentation. Out of the two methods we tried out, one of them is found to be suitable for OCR task. The FGM algorithm provides an average accuracy of 89.97% for Tamil character segmentation.

Keywords
Level Set, Active Contours, Binarization, Segmentation.

1. INTRODUCTION
Binarization involves assigning a value of '0' or '1' with threshold values as mean value of the pixel after the process of converting the same to one bit value, each from the main image. Those pixel values exceeding the threshold are given'1' and the rest are assigned '0'. One of the main process of Binarization, helps separating text pixels from the background. The process of image analysis includes conversion from gray-scale to binary image by pre-processing. Large variability of appearance of texts (font style, size), complex background, occlusions, object shadows, highlights from shiny object parts, and differences of color brightness of objects which make the problem of text binarization on still images as hard. The techniques proposed by J.Ohya et al.[1], Y.Zhong et al.[2].,O.D.Trier et al.[3].,A.T.Abak. et al[4].,M.Sezgin et al.[5]., are usually used for image binarization.

Image Binarization methodologies that exist today are classified into two categories: Thresholding based and grouping based. When the text to be detected is well contrasted with the background most of the existing algorithms perform better, however these algorithms fail when there is no sufficient distinction between background and text to be detected [18]. Niblack's [7] and Sauvola [8] are the most widely used adaptive thresholding algorithms. The methods mentioned are robust for unevenly illuminated and varying colors than those on global ones, but they do suffer regarding the dependency of parametric values. Thresholding based methods use global or local threshold to separate text from background [18]

The global thresholding techniques convert a full image from gray values to two valued (black and white) using one threshold. If the pixel value is greater than the global threshold it is converted to white (background) else it is converted to black (foreground). The determination of this global threshold could be made using histogram properties, busyness properties and global edge information or determined a priori by the user.

Considering the image to gray level histogram I(x,y), which has objects darker than its background, where the object and the background pixels have gray levels which can be classified into two dominant modes. Choosing a threshold `T' we can separate the object from its background. The next is choosing an object point (x,y) which has I(x,y) is greater than T, otherwise the point chosen is called background point. Finding an intensity value separating the two peaks from the histogram is to determine the threshold T in the bimodal case; in normal case the high peak denotes the background (white) and the intermediate peak represents the print (black). The process of estimating the valley between the right peaks of a histogram can be tedious, in case where the histogram does not have a single pure valley between the two peaks.

Simple situations can be dealt with the global thresholding technique, one of the common issues of global thresholding occurs when the gray level transition between the object and the background is gradual and inconsistent across objects in the image. This leads to the use of local thresholding.

Mostly the pixel to be converted is examined in all global binarization methods while determining the subsequent binary value of that pixel. But in addition to this the neighborhood pixels around this pixel is also examined in local thresholding. In case the pixel selected is considerably darker than the neighborhood pixels, it is converted to black else to white. The binary value can be determined with the help of local conditions using the neighborhood information and just not based on the single threshold computed from an image-wide standpoint. The main drawback of the local methods is their high computational cost and their tendency to fail to segment the inner parts of thick characters.

Though there exist a lot of approaches for image binarization of document images, most of these approaches fail for images of complex nature such as printed document images and handwritten images. Level set framework is used to binarize the image in an optimum way, which offers an elegant way to integrate various priors. To our knowledge this work is among the first ones to take advantage of the level set method to achieve good results in document image binarization..

The partitioning of a single digital image into multiple segments refers to segmentation. The aim of segmentation is to make the image that is easily understood by the machine for processing. The segmentation helps identify object and

object boundary like lines, curves, etc., that are present on the image.

In this paper we investigate the use of two recently introduced convex optimization methods, fast global minimization (FGM) algorithm [15] and selective local/global segmentation (SLGS) algorithm [16] for binarization and comparing the performance. The importance of this approach is that it obviates the need for computationally expensive pre-processing tasks like skew correction, line segmentation, word segmentation. The method directly does the binarization and character segmentation. In case the document is highly degraded, preprocessing operation of denoising and deblurring is required.

The paper is organized as follows: Section 2 describes the Theory behind the method, Section 3 describes the recent approaches. In Sect. 4, the experimental results. Finally, the conclusion is provided in Sect. 5.

2. THEORY BEHIND THE METHOD

The method uses optimization theory and level set theory to do segmentation. Image binarization and segmentation can be done using Variational methods. Fast minimization algorithm for general active contour model [14][16] and [17], is used in this paper which is considered as simple and fast while existing techniques require separate methods for binarization and segmentation.

2.1 Level Set Theory

The level set method is a numerical technique used for tracking interfaces and shapes. Level set is optimization method to extract or segment the object by its shape from an image. These interface can have sharp corners .The technique can find a wide range of application including problems in image processing, computer graphics, shape of snowflakes.

Consider an image f with background and foreground. Boundaries can be detected using curve evolution. The boundary of an open domain can be represented using a curve C as the isoline of a Lipschitz continuous function:.

$$f : \Omega \to R$$

$$\phi : \Omega \to C \text{ , where } C = \{(x, y) \in \Omega : \ \phi(x, y) = 0\}$$

For a general curve C, ϕ it can be defined as a signed distance function to C

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

We have to operate on the family of curves C(t) over the time $t \geq 0$ to perform curve evolution.

2.2 Active Contours

Active contour is the process of curve fitting iteratively to an image based on its shapes and the image values until it stabilizes. Two types of active contour curves are used to fit images. They are parametric and geometric active contours. Parametric active contours are used to store vertices and each of vertexes can be moved iteratively. Geometric active contours are stored as coefficients and new coefficients are computed. Geometric active contours have many advantages over parametric active contours, such as computational simplicity and the ability to change curve topology during deformation

2.2.1 Active Contour models

They are based on two criterions namely regularization of the curve and stopping criterion. It can be used for boundary detection. There are two types of models edge based models and region based.

2.2.1.1 Edge based models

This model divides the image based on the discontinuities in image f. An edge function g(x, y) is defined which has a value close to zero near edges and large value away from the edges. It is given by (1).

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

Geodesic active contours, an edge based model proposed by Caseless et al [20], Kichenassamy et al [19] is:

$$\min_{\phi} \left\{ J(\phi) = \int_{\Omega} g(x, y) \mid \nabla H(\phi(x, y) \mid dxdy. \right. \tag{2}$$

2.2.1.2 Region Based Models.

Region based methods segment an image into objects and background. This is done on the basis of similarity of intensity of pixels [16],[17] . Here the following energy is minimized:

$$\begin{aligned} \min_{C,c_1,c_2} E(C,c_1,c_2) = & \int_{inside(C)} (f - c_1)^2 dxdy \\ & + \int_{outside(C)} (f - c_2)^2 dxdy \\ & + \lambda Length(C) \end{aligned} \tag{3}$$

Heaviside function

$$H(\phi) = \begin{cases} 1 \ if \ \phi \geq 0 \\ 0 \ if \ \phi < 0 \end{cases}, \tag{4}$$

and the delta function

$$\delta(\phi) = H'(\phi) \tag{5}$$

in the weak sense.

In the implicit representation the minimization problem can be represented as:

$$\begin{aligned} \min_{f,c_1,c_2} E(C,c_1,c_2) = & \int_{\Omega} (f - c_1)^2 H(f) dxdy \\ & + \int_{\Omega} (f - c_2)^2 (1 - H(f)) dxdy \\ & + \lambda \int_{\Omega} \left|\tilde{N} H(f)\right| \end{aligned} \tag{6}$$

and the inside means c1 and outside mean c2 is obtained as:

$$c_1 = \frac{\int_{\Omega} f(x, y) H(\phi) dxdy}{\int_{\Omega} H(\phi) dxdy} \ , \tag{7}$$

$$c_2 = \frac{\int_{\Omega} f(x, y)(1 - H(\phi)) dxdy}{\int_{\Omega} (1 - H(\phi)) dxdy} \tag{8}$$

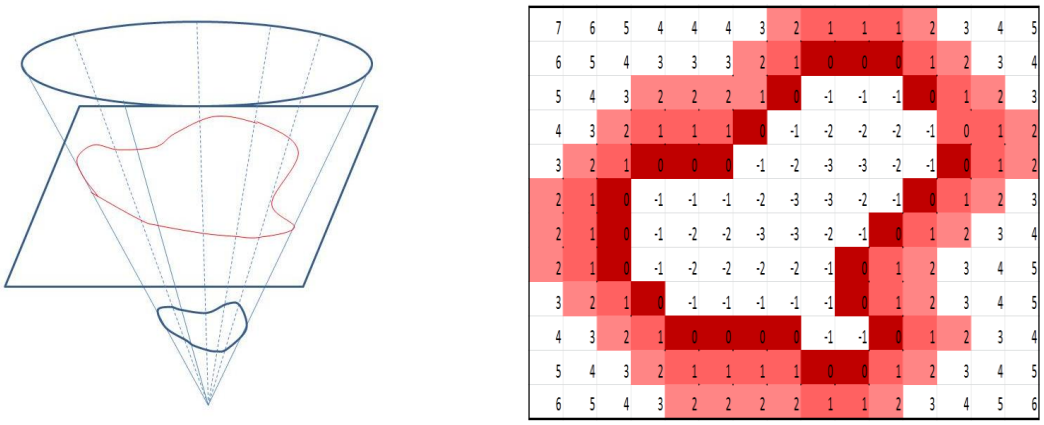

Fig 1: The Level Set Visualization

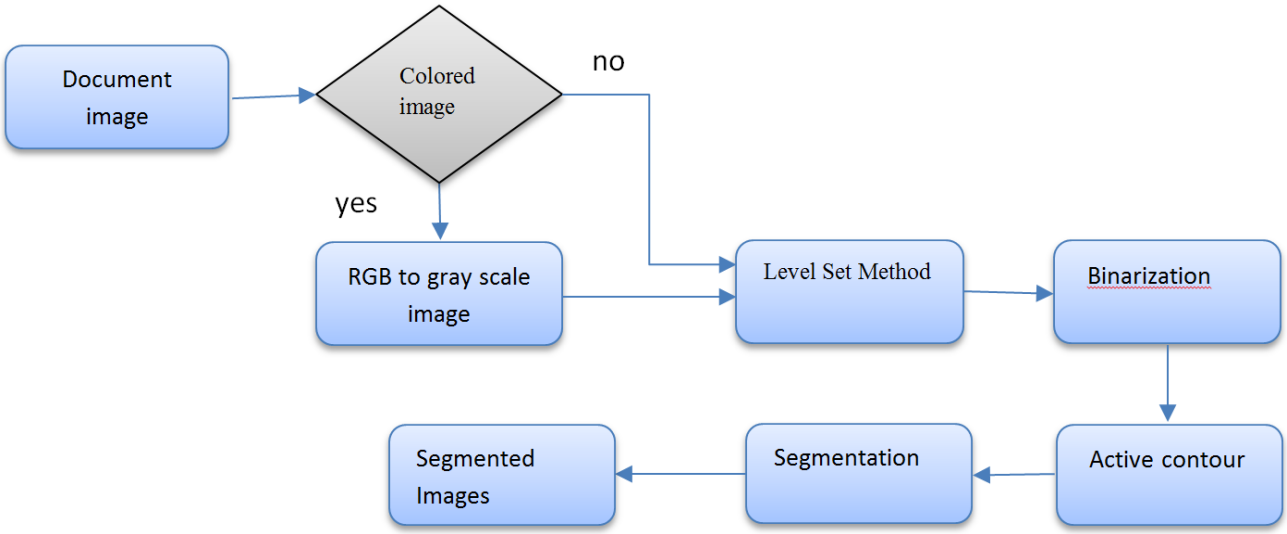

Fig 2: The System Block Diagram of Binarization and Segmentation

For minimization in

$$\phi, \quad \frac{\partial \phi}{\partial t} = -\frac{\partial E}{\partial \phi}$$

is used. Simplification gives:

$$\phi(x, y, 0) = \phi_0(x, y), \tag{9}$$

$$\frac{\partial \phi}{\partial t} = \delta(\phi)\left[-(f - c_1)^2 + (f - c_2)^2 + \lambda K(x, y)\right] \tag{10}$$

Level Set Method can solve (6) but it gives only a local minimizer because its energy is non-convex. This model is less dependent on initial curve but if the initial contour is far away from object it may not extract it. The active contour method tries to convexify the energy and minimize it to compute a global minimizer

3. RECENT APPROACHES

3.1 Block Diagram

For image segmentation and binarization, we make use of Active Contour Method (ACM) and Level set method. Active contour models are used to identify the object boundaries of an image. The contours evolve and fit on to the boundaries of the characters, this helps us easily binarize the document. We have marked the contours with red and given the labels C1, C2, C3, for each contour for better visualization. The outer contour C1 gives the boundary of the character. It is the longest contour. The contours C2 and C3 correspond to the contours drawn for each loop in the character. Contours at the boundaries of the character are equivalent to getting the characters segmented. So by applying the active contour models we can do binarization and segmentation simultaneously.

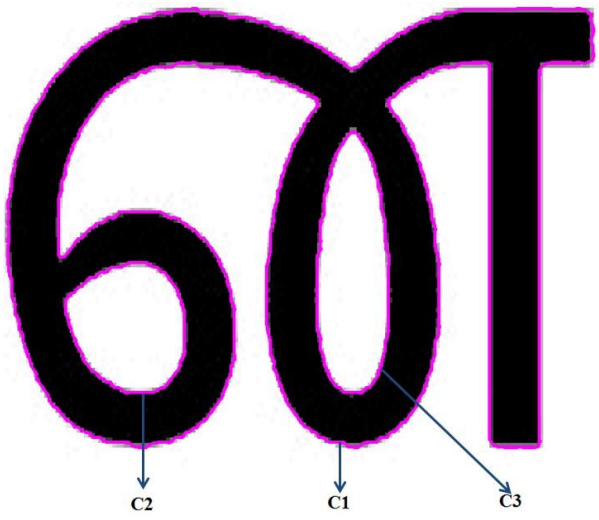

Fig 3: Contour Fitting operation diagram

4. EXPERIMENTS AND RESULT

The present paper apart from experimenting the methods that performs binarization and character segmentation, also compares other methods present in the literature for same kind of problem. The algorithm was tested on document images representing both handwritten and printed documents, SLGS and FGM methods are used for image document analysis. Analysis shows that Fast global minimization algorithm(FGM) is better than the selective local or global segmentation method(SLGS). FGM algorithm works faster while SLGS algorithm takes some time. Fast minimization algorithm can provide more accuracy in binarization as well as in segmentation

4.1 Selective local or global segmentation algorithm

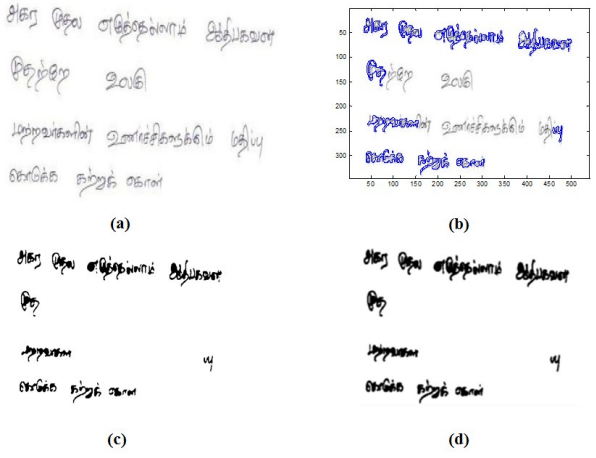

Fig 4: Selective local or global segmentation operations explained. (a) represents the original image. (b) Curve fitting in Active Contour. (c) Image obtained after applying Thresholding. (d) Binarized Image

4.2 Fast Global minimization algorithm

Fast Global minimization algorithm using Level set

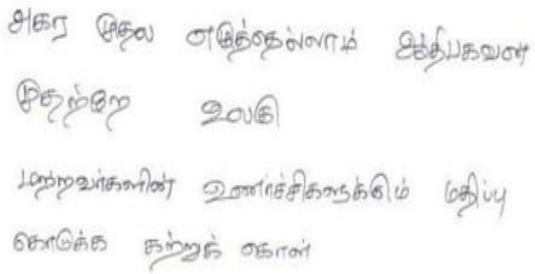

Fig 5: Original Image

method is best algorithm when compared with Global minimization algorithm. When we compared with few documents Fast Global minimization algorithm provides the binarized document, approximately 22 times faster than the previous method, so we have selected Fast global minimization algorithm which performs faster computing. The performance comparison of Fast Global Minimization algorithm is shown in Table:1

Fig 6: Contour evolved Image

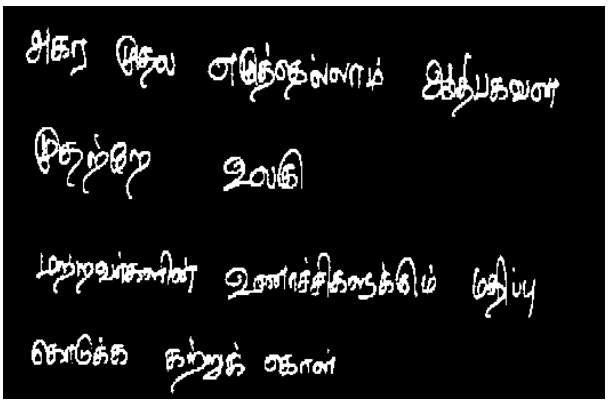

Fig 7: Binarized Image

Table 1. Performance comparison of SLGS and FGS algorithms in Seconds

Binarization	SLGS algorithm	FGM algorithm
Character less than 10	2.886	1.2636
Character less than 50	77.0645	1.7165
Character less than 100	430.0948	3.978
Character less than 500	562.76	11.654
Character more than 1000	784.295	24.7886

4.2.1 Image Segmentation

Level Set algorithm returns contours of each object. That is in MATLAB, the contour command when applied to the output of Level Set method gives the contour of each letter there by allowing character segmentation.

The algorithm is applied to image documents and accuracy is computed. The inaccuracy mainly arises due to touching of characters.

To improve the accuracy we need to do the following:

Fig 8: Input Image - Active contour evolved

Fig 9: Binarized Image

Fig 10 : Segmented Image

- Identification of touching characters, If contour length is more it may be due to touching characters.
- Incorporating Spell checkers.
- Dictionary

The Time taken for binarization and segmentation is comparatively less when compared with Global minimization using thresholding algorithm.

The result is compared for different sets of varying characters with their correctness and accuracy is as follows: Table 2.

Table 2. Result Analysis

Segmentation	correct	Incorrect	percentage accuracy	time (s)
6 characters	6	0	100	0.36
26 character	23	3	88.46	1.5912
55 character	49	6	89.09	3.666
88 character	78	10	88.63	6.24
156 character	139	17	89.11	11.32
253 character	214	39	84.58	19.56

5. CONCLUSION

We have experimented Active contour based segmentation and binarization. To detect the boundaries of the object we used energy minimization procedure and obtained well segmented objects. Binarization converts the document image into binary image and also helps to remove the noise.

6. REFERENCES

[1] J. Ohya, A. Shio, and S. Akamatsu. "Recognizing characters in scene images". IEEETrans. Pattern, Anal.Mach.Intell., 16(2), 1994, pp.214-220.

[2] Y. Zhong, K. Karu, and A.K. Jain. "Locating text in complex color images." , *Proc. of 3rd Int. Conf. Document Analysis and Recognition*, 1995, 146 - 149 vol.1.

[3] 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.

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

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

[6] PavlosStathis, ErginaKavallieratou, Nikos Papamarkos, "An Evaluation Technique for Binarization Algorithms", Journal of Universal Computer Science, vol. 14, no. 18 ,2008, pp. 3011-3030.

[7] W. Niblack, "An Introduction to Image Processing", Prentice-Hall, En- glewood Cliffs, NJ 1986 , pp. 115116.

[8] J. Sauvola and M. Pietaksinen, "Adaptive document image binarization," Pattern Recogn. 33, 2000 , pp. 225236.

[9] J. N. Kapur, P. K. Sahoo, and A. K. C. Wong, "A New Method for Gray-Level Picture Thresholding Using the

Entropy of the Histogram," Computer Vision, Graphics and Image Processing 29, Mar. 1985, pp.273-285.

[10] P. K. Loo and C. L. Tan. "Adaptive Region Growing Color Segmentation for Text Using Irregular Pyramid". Document Analysis Systems VI Lecture Notes in Computer Science, Volume 3163/2004, 2004, pp. 103-106, DOI: 10.1007/978-3-540-28640-0-25.

[11] C. Fung, R. Chamchong, "A Review of Evaluation of Optimal Binarization Technique for Character Segmentation in Historical Manuscripts,"3rd Int. Conf. Knowledge Discovery and Data Mining, 2010 ,pp.236-240.

[12] N. Dinh, J. Park, G. Lee, "Korean Text Detection and Binarization in Color Signboards", Int. Conf. ALPIT ,2008, pp. 235-240.

[13] Stathis, E. Kavallieratou, N. Papamarkos, "An evaluation survey of binarization algorithms on historical documents", 19th ICPR, 2008, pp.1-4

[14] X. Bresson, S. Esedo, P. Vandergheynst, J.-philippe Thiran, and S. Osher, "Fast Global Minimization of the Active Contour / Snake Model," Journal of Mathematical Imaging and Vision, 2007 28: 151167.

[15] X. Bresson, "A Short Guide on a Fast Global Minimization Algorithm for Active Contour Models.," Energy, pp. 1-19, 2009.

[16] Zhang, K., Zhang, L., Song, H., and Zhou, W. (2010). "Active contours with selective local or global segmentation : A new formulation and level set method.", Image and Vision Computing 28(4), 668-676. Elsevier B.V. doi: 10.1016/j.imavis.2009.10.009.

[17] David Rivest-Hnault , Reza Farrahi Moghaddam ,Mohamed Cheriet "A local linear level set method for the binarization of degraded historical document images", Springer-Verlag 2011

[18] Basura Fernando, Sezer Karaoglu, Alain Trmeau, "Extreme Value The- ory Based Text Binarization In Documents and Natural Scenes", 3rd Int. Conf. Machine Vision, Hong Kong, 2010.

[19] Kichenassamy, S., Kumar, A., Olver, P., Tannenbaum, A., and Yezzi,A. " Gradient flows and geometric active contour models", Proc. ICCV, Cambridge. 1995.

[20] Caselles, V., Catte, F., Coll, T., and Dibos, F. "A geometric model for active contours", Numerische Mathematik, 66:131. 1993.