Combined Method of Level set with impact on Pre-processing for binarization of document images in Tamil Script

Asha Ashok, Dhivya S, Jansirani S, K.P. Soman
Centre for Excellence in Computational Engineering and Networking
Amrita Vishwa Vidyapeetham, Coimbatore.

ABSTRACT

Binarization and Segmentation are considered to be the vital tasks in Optical Character Recognition (OCR) for document digitization. This paper discusses the applications of powerful Level set methodologies on these important tasks associated with OCR. Results acquired for these essential procedures in turn govern the accuracy of the OCR system. Conventionally Otsu method and Histogram profiling methods were used for binarization and segmentation purpose [1]. In this paper, we try to replace these different methods by a single procedure based on Level Set methodology, where segmentation and binarization for any document could be efficiently done in a single step. The main advantage of this method is that it does not require separate paragraph segmentation, line segmentation and character segmentation. We have also compared the working and performance of two algorithms based on Active Contour or Level Set Model: Selective Binary and Gaussian Filtering Regularized Level Set (SBGFRLS) by Zhang [2] and Split Bresson's Chan Vese Level Set methods by Bresson [3] in the case of real data and test data embedded with noise taken from Tamil script.

General Terms

Binarization, Segmentation, Optical Character Recognition, Document Digitization.

Keywords

Level Set Model, Contour, Edge based Model, Region Based Model, Image Segmentation, Parametric Active Contour, Non Parametric Active Contour.

1. INTRODUCTION

Tamil is a Dravidian language which has acclaimed one of the World Classical Language status. In today's world, the usage of paper is declining at a faster rate with the advent of increased usage of computers. It is due to the fact that computers are incredibly fast and the cost of computation and storage becomes cheaper with time. Hence, a mechanism is inevitably needed which helps to bring all sorts of information on the digital platform. This proves the fact that Document Image binarization has an important role to play as far as document-digitization is considered [4-6]. This paper is basically written with an aim to provide a detailed study about document binarization and image segmentation, mainly carried out on Tamil documents. Once, this is done efficiently, then the remaining tasks related with pattern-classification, automated text classification, information retrieval systems, natural language processing etc., are carried out much more easily with less consumption of time. But challenges associated with old and historic documents are unlimited due to the fact that it may include several defects in their background and on character edges. These defects originate from scanning, paper aging and spreading of written ink, presence of stains or of holes etc. In spite of such problems, the pre-processing must be carried out efficiently to increase the accuracy of the system. Level Set methods are used to segment or extract the object by its shape from an image. Our goal is to better understand the pros and cons of two powerful level set methods proposed by Zhang [2] and Bresson [3] applied in Tamil script.

1.1 Requirement for Digitizing Documents

A large community of people thinks it is mandatory to preserve the historic paper based documents. Moreover, many of the historic documents are degraded. The reasons for degrading includes, ageing, human manipulation, ink fading, folding marks, holes and spots, bleed through ink, paper appearance, its texture, non-uniformity of stroke pixel intensities, due to ink smearing and variations in pen pressure, certain font shapes with very thin segments and scanning process can lead to uneven illumination in the image. So, the need arises that the aged old documents must be preserved. At the same time, it is not only preserved, but also transformed to a form which is quite easily accessible, editable, and also at the same time portable. At present, all the work carried out is done to digitize the historic paper based documents. This demand in digital copies has resulted in an increased demand for journals, articles and other literature to be available in a digitized version. Thus, the first step in converting a paper based document into digital form is to scan the document using a high resolution scanner. Once the scanned image of the required document is obtained, then an effort is made to transform it into a machine editable text. This is achieved by applying the method of Optical Character Recognition [7]. Binarization and segmentation are the most important pre-processing steps carried out as far as any OCR work is considered [8].

1.2 Need for Binarization

The complete color representation of an image is not required when character recognition tasks are concerned, and so pages are often scanned or converted to a greyscale (8bpp) or bi-level (1bpp) color depth. In grayscale, each pixel represents one of 256 shades of gray, and in a bi-level image each pixel is assigned one of two values representing black or white. While both of these methods will allow a digital image to be stored in a smaller space (fewer bpp), they can suffer from information loss as the original color value is more coarsely approximated. The process of converting a color or grayscale image to bi-level format is referred to as binarization. Several approaches to binarization have been discussed in the literature but they typically fall into one of two categories [8].

Global methods treat each pixel independently, converting each to black or white based on a single threshold value. These are usually not suitable for degraded document image, because these images do not have a clear bimodal pattern that separates foreground and background. But in clear document images, where there is a suitable dissimilarity between fore ground and back ground, global methods do work well. If pixel's color intensity is higher than the global threshold it is assigned one value, otherwise it is assigned the opposite value. In contrast local methods, make use of the color information in nearby pixels to determine an appropriate threshold for a particular pixel. Typically the two colors used for a binary image are black and white though any two colors can be used. The simplest way to use image binarization is to choose a threshold value, and classify all pixels with values above this threshold as white, and all other pixels as black. The problem lies in selecting the precise threshold. This still remains as an unsolved problem due to different types of document degradations, image contrast variations, bleeding-through, and smear. Other factors which make binarization a difficult task is the presence of strong noise, complex patterns and/or variable modalities in gray-scale histogram.

1.3 Pre-requisite to segment Images

Image segmentation [9] is one of the most crucial steps in the image analysis. The main goal is to divide an image to some parts, which connect intensely by objects of authenticity. Many universal and effective segmentation methods can be written as variational/PDE based models [10]. This category of variational models has revealed to be very effective in many applications, specifically in the handling and study of medical images [11-13]. While there are many approaches to image segmentation, this paper will focus on recently proposed methods which can be cast in the form of totally convex optimization problems. This convexity property allows segmentations to be computed using fast elliptic solvers.

1.3.1 Solutions for Segmentation Problem

Several different variational frameworks for image segmentation have been proposed. Most of which fall into one of two categories of a novel Active Contour Model formulation. An active contour is an energy minimizing spline that detects specified features within an image [22]. It is a flexible curve (or surface) which can be dynamically adapted to required edges or objects in the image (it can be used to automatic objects segmentation). The main such group is the geodesic active contour (GAC) / snakes model/edge based model [14]. The second approach to segmentation that we shall consider is a technique based on the Mumford Shah model/region based model [15].

1.3.2 Various Segmentation Models

Basically, there are two methods commonly used in Image segmentation. These are edge-based and region-based models. They are classified depending upon how image information is utilized to match the active contour with the object under consideration.

1.3.2.1 Edge Based Models

They are also referred as Geodesic Active Contour or Snakes Model. It is based on discontinuities in the intensity of an image. Partitioning of an image happens due to the event of unexpected changes in intensity. Information pertaining to image gradients is utilized to identify the edges of objects in an image

1.3.2.2 Region Based Models

It is also commonly referred as Active Contour without Edge Model. It looks for uniformity or similarity based on any properties regarding color, texture or even intensity within a particular region. They are generally robust to noise but the main drawback happens when the object and the surrounding background have similar characteristics. Image segmentation hence becomes a challenging problem for region based Models. Figure 1, shows how an initial contour finally matches with the image data corresponding to a triangle.

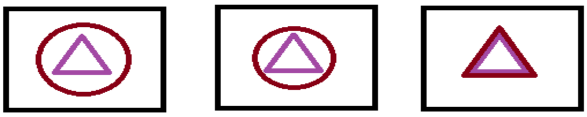


Fig 1: A smooth curve which matches to Image data.

1.4 Basic Notion of Active Contour

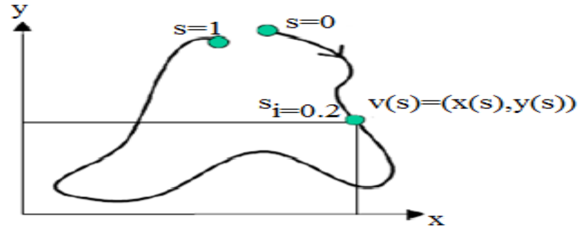


Fig 2: Basic form of Active Contour

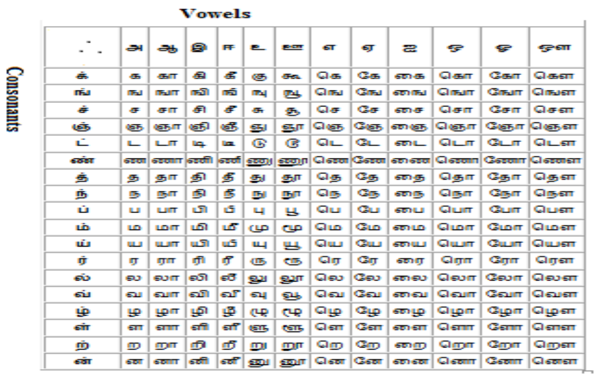
The starting position of the contour is stated by the user initially and then the contour is moved by image driven forces to the boundaries of the desired objects. The contour is defined by the curve *v(s)*, which in turn is a function of *(x(s),y(s))*. *s* is a value ranging from 0 to 1.It minimizes the energy *E*, which is comprised of internal energy, E_{int} and external energy, E_{ext} .The internal forces always try to contract the contour. Whereas the external forces which always depend on content of the image, prevents the contour from propagating outward and thus helps in reaching the model to the precise object boundary. Figure 2 illustrates the representation of a contour.

1.4.1 Alternate Classes of Active Contour Models

Another kind of classification partitions the active contour into parametric and non-parametric active contour. Basically, two forms of deformable models exist. In the parametric form, also known as snakes, an explicit parametric depiction of the curve is used. This form is not only dense, but is robust to both image noise and boundary gaps as it compels the extracted boundaries to be smooth. However, the performance is strictly affected when deformation involves splitting or integration of parts. In contrast, the implicit deformable models, also called implicit active contours or level sets, are intended to handle topological changes naturally. However, unlike the parametric form, they are not robust to boundary gaps and suffer from other shortages also [16].

1.5 Insight into Tamil language

Tamil is also spoken in countries like Sri Lanka, Singapore, Malaysia and Mauritius. Tamil employs agglutinative grammar, where suffixes are used to mark noun class, number, and case, verb tense and other grammatical categories [17]. Tamil has 12 vowels and about 18 consonants. These vowels and consonants are combined together to produce 216 composite characters and 1 special character yielding a total of (12+18+216+1) 247 characters. Figure 3 displays the characters of Tamil alphabets.



Vowels

Consonants

ஃ	அ	ஆ	இ	ஈ	உ	ஊ	எ	ஏ	ஐ	ஒ	ஓ	ஔ
க்	க	கா	கி	கீ	கு	கூ	கெ	கே	கை	கொ	கோ	கௌ
ங்	ங	ஙா	ஙி	ஙீ	ஙு	ஙூ	ஙெ	ஙே	ஙை	ஙொ	ஙோ	ஙௌ
ச்	ச	சா	சி	சீ	சு	சூ	செ	சே	சை	சொ	சோ	சௌ
ஞ்	ஞ	ஞா	ஞி	ஞீ	ஞு	ஞூ	ஞெ	ஞே	ஞை	ஞொ	ஞோ	ஞௌ
ட்	ட	டா	டி	டீ	டு	டூ	டெ	டே	டை	டொ	டோ	டௌ
ண்	ண	ணா	ணி	ணீ	ணு	ணூ	ணெ	ணே	ணை	ணொ	ணோ	ணௌ
த்	த	தா	தி	தீ	து	தூ	தெ	தே	தை	தொ	தோ	தௌ
ந்	ந	நா	நி	நீ	நு	நூ	நெ	நே	நை	நொ	நோ	நௌ
ப்	ப	பா	பி	பீ	பு	பூ	பெ	பே	பை	பொ	போ	பௌ
ம்	ம	மா	மி	மீ	மு	மூ	மெ	மே	மை	மொ	மோ	மௌ
ய்	ய	யா	யி	யீ	யு	யூ	யெ	யே	யை	யொ	யோ	யௌ
ர்	ர	ரா	ரி	ரீ	ரு	ரூ	ரெ	ரே	ரை	ரொ	ரோ	ரௌ
ல்	ல	லா	லி	லீ	லு	லூ	லெ	லே	லை	லொ	லோ	லௌ
வ்	வ	வா	வி	வீ	வு	வூ	வெ	வே	வை	வொ	வோ	வௌ
ழ்	ழ	ழா	ழி	ழீ	ழு	ழூ	ழெ	ழே	ழை	ழொ	ழோ	ழௌ
ள்	ள	ளா	ளி	ளீ	ளு	ளூ	ளெ	ளே	ளை	ளொ	ளோ	ளௌ
ற்	ற	றா	றி	றீ	று	றூ	றெ	றே	றை	றொ	றோ	றௌ
ன்	ன	னா	னி	னீ	னு	னூ	னெ	னே	னை	னொ	னோ	னௌ

Fig 3: Tamil characters.

The motivation behind this paper is to conduct several experiments to understand which level set algorithm proposed by Zhang or that of Bresson gives a better segmented and binarized output when an input of Tamil script is specified.

2. LEVEL SET MODEL FOR IMAGE BINARIZATION

Assume *C* to be a closed curve or a family of curves, with parametrizations [0 ,1] chosen in such a way that *C(0)=C(1)*. The main aim is to evolve the curve *C* till the boundary of the desired object is detected[18]. An isoline function in the form *L(x,y,t)=0* is assumed for edge detection by evolving the curve. Here *L* is the zeroth level set function. Level Set is nothing but a set of points for which the level set function is a constant value. For curve evolution to detect object boundaries, we divide the image into two regions: Area enclosed within the curve *C* and area outside of the curve *C*. For this representation, a signed distance function is utilised. It is represented mathematically in Equation 1 and the geometrical interpretation is depicted in Figure 4.

$$\phi(x,y)=\begin{cases} dist\left((x,y),C\right) & \text{if } (x,y)\in \text{ inside } C \\ -dist\left((x,y),C\right) & \text{if } (x,y)\in \text{ outside } C \\ 0 & \text{if } (x,y)\in C \end{cases} \qquad (1)$$

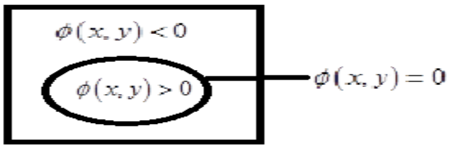


Fig 4: Geometrical Representation of Signed distance function.

The curve, C is progressed such that the level set converges correctly at the object boundary. For this, the curve is evolved over time, $t \geq 0$ which are isolines of functions $\phi(x,y,t)$ such that, $C=\{(x,y)\in\Omega:\phi(x,y,t)=0\}$ (2)

The curve moves in normal direction based on the equation

$$C'(t)=F\overrightarrow{N} \qquad (3)$$

F is the speed function based on some criteria which leads to the stopping of the curve at the required boundary. $\overrightarrow{N}$ is the unit normal vector to curve *C*. Assuming *C(t)* is implicitly represented by ϕ and differentiating this with respect to time *t*, finally it is obtained as,

$$\frac{\partial\phi}{\partial t}=F\left|\nabla\phi\right| \qquad (4)$$

2.1 Design of Euler Lagrangian Equation in Level Set Methodology

Euler Lagrangian (EL) equation gives solution to obtain the maxima or minima of a functional of the form,

$$I\left[y(x)\right]=\int_{x_1}^{x_2}F\left(x,y(x),y'(x)\right)dx \qquad (5)$$

It implies that if J is defined by a functional, which is a function of a function, in the following form,

$$J=I\left[y(x)\right]=\int_{x_1}^{x_2}F\left(x,y(x),y'(x)\right)dx \qquad (6)$$

then J has a stationary value if the EL differential equation $\frac{\partial F}{\partial y}-\frac{d}{dx}\left(\frac{\partial F}{\partial y'}\right)=0$ is satisfied.

Similarly for the functional of form, $\int F\left(x,y,\phi(x,y)\right)$ EL solution is formulated as,

$$\frac{\partial F}{\partial\phi}-\left(\frac{\partial}{\partial x}\frac{\partial F}{\partial\phi_x}+\frac{\partial}{\partial y}\frac{\partial F}{\partial\phi_y}\right)=0 \qquad (7)$$

For curves to evolve and settle on object boundaries, an expression is found for energy measure $E(\phi)$, which is then minimized with respect to ϕ. This is equivalent as,

$$\frac{\partial\phi}{\partial t}=-\left[\frac{\partial F}{\partial\phi}-\left(\frac{\partial}{\partial x}\frac{\partial F}{\partial\phi_x}+\frac{\partial}{\partial y}\frac{\partial F}{\partial\phi_y}\right)\right] \qquad (8)$$

This method is known as time-margin initialization for curve evolution.

2.2 Edge Detection by Active Contour Models

As mentioned earlier for the case of image segmentation models, similarly we can classify the models used for image binarization by Active contour methodology into two. It is known as the Edge based and Region based models.

2.2.1 *Deep Understanding of Edge based Models*

An edge-stopping function given by $g\left(x,y\right)$ is defined as,

$$g\left(x,y\right)=\frac{1}{1+\left|\nabla\left(G_\sigma * f\left(x,y\right)\right)\right|^2} \qquad (9)$$

Where G_σ is the Gaussian filter, obtained by sampling the Gaussian distribution with standard deviation σ .This edge stopping function has a small value near to the edges (due to the large value in gradient). This accounts for the converging of the active contour exactly at the object boundary.

2.2.2 Deep Understanding of Region based Models

The main theory is that they partition an image into sub-regions based on similarity of intensity between pixels. The energy to be minimized is given below,

$$\min_{C,C_1,C_2} E\left(C,C_1,C_2\right)=\int_{inside(c)}\left(f-C_1\right)^2 dxdy \\ +\int_{outside(c)}\left(f-C_2\right)^2 dxdy+\lambda Length\left(C\right) \qquad (10)$$

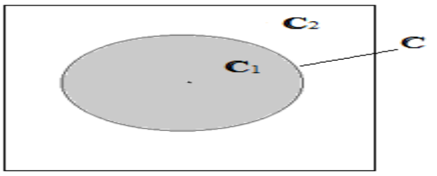


Fig 5: Splitting of an image by Region based model into C$_1$&C$_2$

Figure 10 shows that the average pixel value inside the curve and outside the curve *C* is *C$_1$* and *C$_2$* respectively. λ is the regularization parameter. A Heaviside function is defined

$$H\left(\phi\right)=\begin{cases}1 \text{ if } \phi\geq 0\\ 0 \text{ if } \phi<0\end{cases} \qquad (11)$$

It is assumed delta function $\delta\left(\phi\right)=H^{'}\left(\phi\right)$ in weak sense. While representing in implicit form, the minimization problem is casted as,

$$\min_{\phi,C_1,C_2} E(C,C_1,C_2)=\int_\Omega (f-C_1)^2 H(\phi)dxdy \\ +\int_\Omega (f-C_2)^2(1-H(\phi))dxdy+\lambda\int_\Omega\left|\nabla H(\phi)\right| \qquad (12)$$

Intensities *C$_1$* and *C$_2$* is obtained as

$$C_1(\phi)=\frac{\int_\Omega f\left(x,y\right)H\left(\phi\left(x,y\right)\right)dxdy}{\int_\Omega H\left(\phi\left(x,y\right)\right)dxdy}$$

$$C_2\left(\phi\right)=\frac{\int_\Omega f\left(x,y\right)\left(1-H\left(\phi\left(x,y\right)\right)\right)dxdy}{\int_\Omega\left(1-H\left(\phi\left(x,y\right)\right)\right)dxdy} \qquad (13)$$

Minimization in ϕ and substituting for *C$_1$* and *C$_2$*, we obtain,

$$\phi\left(x,y,0\right)=\phi_0\left(x,y\right), \\ \frac{\partial\phi}{\partial t}=\delta\left(\phi\right)\left[-\left(f-c_1\right)^2+\left(f-c_2\right)^2+\lambda K\left(x,y\right)\right] \qquad (14)$$

$K\left(x,y\right)$ is the curvature, mathematically given as,

$$K\left(x,y\right)=\frac{\phi_{xx}\phi_y^2-2\phi_x\phi_y\phi_{xy}+\phi_{yy}\phi_x^2}{\left(\phi_x^2+\phi_y^2\right)^{3/2}} \qquad (15)$$

3. CONCEPTOF PROPOSED METHODS

3.1 Active Contours with Selective Local or Global Segmentation

This model was proposed by Kaihua Zhang and it exploits a combination of edge based and region based model [2]. In this paper, the elementary equation determining curve evolution is

$$\frac{\partial\phi}{\partial t}=spf\left(I\left(x\right)\right)\alpha\left|\nabla\phi\right| \qquad (16)$$

Equation (15) is a modified form of equation (4) where spf is signed pressure function and α is the speed function respectively. spf function assumes values in the range [-1,1] and mathematical interpretation is defined as,

$$spf\left(I\left(x\right)\right)=\frac{I\left(x\right)-\frac{\left(c1+c2\right)}{2}}{\max\left(\left|I\left(x\right)-\frac{\left(c1+c2\right)}{2}\right|\right)} \qquad (17)$$

The algorithm is stated as follows:

a) Initialize level set function as

$$\phi\left(x,t=0\right)=\begin{cases}-\delta \text{ if } x\in\Omega_0-\partial\Omega_0\\ 0 \quad \text{if } x\in\Omega_0\\ +\delta \text{ if } x\in\Omega_0+\partial\Omega_0\end{cases} \qquad (18)$$

Where $\delta>0$, $\Omega_0\subset\Omega$ and $\partial\Omega_0$ is a boundary of Ω_0 .

b) Calculate $\mathrm{C}_1\left(\phi\right)$ and $\mathrm{C}_2\left(\phi\right)$ as in equation (13).

c) Evolve level set function based on equation (16).

d) Regularize level set function with a Gaussian filter according to, $\phi=\phi*G_\sigma$

e) Check if the level set function has converged. If converged, stop. Else, return to (b).

3.2 Fast Global Minimization Algorithm for Active Contour Models

This is another variation of Active contour model and was proposed by Xavier Bresson [3]. It is slightly different from the formulation proposed by Zhang's methodology since the evolution of curve happens with an objective to attain additional information from the midpoint of the image. Global minimizer of the active contour model is formulated as

$$\min_c\{F_{AC}(C)=\int_C g_b(C,s)ds+\lambda\int_{C_{in}} g_r^{in}(C_{in},x)dx+\lambda\int_{C_{out}} g_r^{out}(C_{out},x)dx\} \quad (19)$$

C is the contour, C_{in} and C_{out} represents the region inside and outside of C. g_r^{in} and g_r^{out} signifies the inside and outside region where as g_b indicates the boundary region. Moreover, we have

$$\begin{aligned} g_r^{in}(C_{in},x) &= \left(\mu_{in}\left(C_{in}\right)-1\right)^2 \\ g_r^{out}(C_{out},x) &= \left(\mu_{out}\left(C_{out}\right)-1\right)^2 \end{aligned} \tag{20}$$

μ_{in} and μ_{out} are clear as in equation (13). Gradient flow of C which is used to control curve growth is expressed as,

$$\partial_t C = g_b(C,s)k(C,s) + <\nabla g_b, N(C,s)> + h_r^{in}(C_{in},s) - h_r^{out}(C_{out},s) \tag{21}$$

The Heaviside function is given by h_r . By applying Level Set Method, the formulation obtained is,

$$\min_{\phi}\left\{F_{LSM}\left(\phi\right) = \int_{\Omega} g_b\left|\nabla H\left(\phi\right)\right| + \lambda\int_{\Omega} h_r^{in} H\left(\phi\right) + \lambda\int_{\Omega} h_r^{out}\left(1-H\left(\phi\right)\right)dx\right\} \tag{22}$$

But when solving this problem, only a local minimizer is obtained as the optimization problem is non-convex in nature. Hence, the above formulation is convexified to obtain a global minimizer. The final formulation will become,

$$\min_{\phi}\left\{F_G\left(u\right) = \int_{\Omega} g_b\left(x\right)\left|\nabla u\right| + \lambda\int_{\Omega} h_r^{in}\left(x\right)u + \lambda\int_{\Omega} h_r^{out}\left(x\right)\left(1-u\right)dx\right\} \tag{23}$$

The energy functional F_G has turned into convex optimization problem. This is then solved using Bregman iteration. The algorithm for Bresson's Active Contour model is stated as,

a) Fix u , given by contour C initially and update h_r

b) Fix Heaviside function h_r and update u with the iteration

$$\begin{aligned} \left(u^{k+1}, d^{k+1}\right) &= \arg\min_{u\in[0,1]}\left(\int\left|d\right| g_b + \lambda h_r u + \frac{\mu}{2}\left|d - \nabla u - b_k\right|^2 dx\right) \\ b_{k+1} &= b_k + \nabla u_{k+1} - d^{k+1} \end{aligned} \tag{24}$$

c) The last contour is specified by the boundary as,

$$\left\{x \in \Omega \middle| u^{final}(x) > 0.5\right\} \tag{25}$$

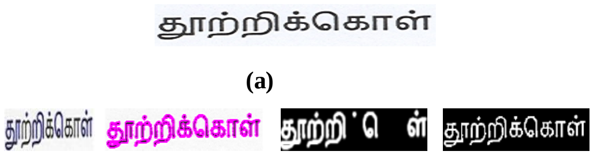
d) Repeat the above steps until level set touches the object boundary.

4. EXPERIMENTS AND DISCUSSION OF RESULTS

The level set algorithms discussed in section 3.1 and 3.2 were used for experimenting with a database comprising of various scanned text documents in Tamil. Numerous kinds of Tamil text including scanned news-papers, magazines, words, damaged documents and also hand-written documents were served as input to both algorithms. The two level set algorithms were applied on all sets of Tamil document images and their performance is compared.

4.1 Experiment with Clean Data

A clean Tamil word, as shown in Figure 6(a), is taken and both Bresson's and Zhang's level set methods were applied on it. The contours drawn by Zhang's and Bresson's algorithm are shown respectively in Figure 6(b) and 6(c). Similarly, in Figure 6(d) and 6(e) displays the result after binarization by Zhang's and Bresson's method. The Zhang's algorithm converges only at the edges of some character of the original text. As a result of which, the binarized result of Zhang's method contains only those particular character for which the contour converged and the remaining portion got cut off. But by Bresson's method, the contour is able to converge properly at the boundary of each particular character, thus leading to a better and full binarized output. For the same input, we were able to notice that Zhang's algorithm gave the binarized result in a time gap of 354 seconds whereas the Bresson's algorithm worked only for about 0.0060 seconds.



(a)

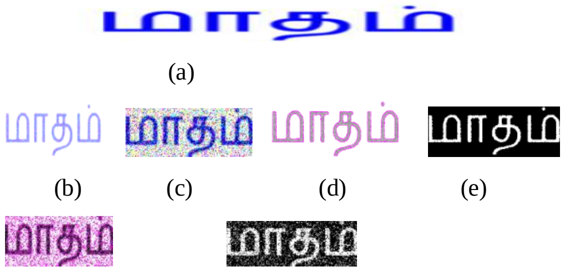
(b) (c) (d) (e)

Fig 6: Results of Zhang's and Bresson's level set method for binarization (a) Original data (b) Contour by Zhang's method (c) Contour by Bresson's method (d) Binarization by Zhang's method (e) Binarization by Bresson's method

The in-built 'contour' function of MATLAB extracts information regarding each contour which converges at the character boundary. It actually returns a contour matrix which contains data about the coordinate positions for all the contours. This contour matrix is 2*N in size. The 2*N matrix obtained contains all the *x* and *y* data along the various contours positioned. By extracting the data for each contour distinctly and summing the distance between corresponding successive-pairs, we are able to calculate the length of each contour.

4.2 Experiment with Noisy Data

The next experiment includes the level set algorithms applied on data surrounded with noise. It is possible in MATLAB by 'imnoise' function. With this, we are able to embed the image with 'Gaussian', 'salt & pepper' and 'poisson' variety of noises. In the former experiment, a clean word is chosen and is then 'imnoised' with Gaussian noise of intensity 50%. This is shown respectively in Figure 7(a) and 7(b). The Zhang's level set algorithm was not able to converge at the edge of any character even after running for about 100 iterations. But the level set algorithm by Bresson gave a complete but blurred output due to the presence of high intensity noise. The result was obtained for Bresson's method in 0.05 seconds with just 4 iterations. The segmented and corresponding binarized result of Bresson's level set methodology is shown in Figure 7(d) and 7(e). The same input in Figure 7(a) is 'imnoised' with 'salt & pepper' noise of intensity 20%. This noisy image is displayed in Figure 7(c). In 'salt and pepper' or 'impulse' noise model, pixels are corrupted randomly by two values, 0 and 255, for an 8-bit image. It may be 'minimum' and 'maximum' gray-level intensity values. When this noisy image is applied to Bresson's method, then flaw is that even the noise is misinterpreted to be part of character and contour is also obtained for each noise component. This effect is reflected in the segmented and binarized output of Bresson's method in Figure 7(f) and 7(g). But the result displayed in Figure 7(g) can be enhanced if we include a pre-processing step like diffusion etc.

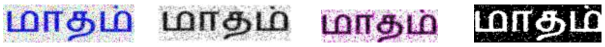


(a)
(b) (c) (d) (e)
(f) (g)

Fig 7: Results of Bresson's level set method for binarization of noisy data (a) Original data (b) Original data in (a) embedded with Gaussian noise of 60% intensity(c) Original data in (a) embedded with impulse noise of 20% intensity (d) Contour drawn by Bresson's method for 'Gaussian' noise image (e) Result for binarization by Bresson's method for 'Gaussian' noise image (f) Contour drawn by Bresson's method for 'impulse' noise (g) Result for binarization by Bresson's method for 'Impulse' noise.

4.3 Prominence of Pre-processing step

The efficiency of the output obtained for the previous experiment is significantly improved with the help of a pre-processing step like diffusion. Usually for degraded and noisy documents, a pre-processing stage is crucial for the dismissal of noisy areas. In this experiment, we have demonstrated the use of 'Coherence –edge diffusion' as a vital pre-processing step to perform high degree of noise-reduction of the image.



(a) (b) (c) (d)

Fig 8: Sample outputs of Bresson's level set method for binarization after a pre-processing step is applied on noisy data (a) Original data with 20% 'impulse' noise as in the previous experiment (b) when 'Coherence Edge diffusion' applied on noisy data (c) Sample output for segmentation by Bresson's level set method (d) Sample binarized output.

Most of the PDE based methods are very crucial for denoising and edge protective applications in images. While Perona and Malik types of nonlinear diffusion are isotropic, an advanced diffusion scheme was introduced by Weickert, which was capable of selecting preferred diffusion directions. This allowed Weickert to define coherence enhancing diffusion where smoothing is directed to be along the image 'isophotes'. The speed of diffusion is controlled by means of a diffusion tensor rather than an absolute value of the gradient of the image intensity just as in second order PDE based anisotropic diffusion methods [19-21].

4.4 Realization of Character Segmentation and Line Segmentation

This experiment demonstrates the result of character and line segmentation by Bresson's Active Contour Method. A sentence is entered in Figure 9(a) and the result in Figure 9(b) shows how the active contour for Bresson's algorithm converges exactly at the boundary of each character. Hence, by the application of Bresson's algorithm, each letter is retrieved and binarized properly. Similarly, the output for line segmentation for an input paragraph comprising of 3-4 lines in Figure 9(c) is also portrayed in Figure 9(d). One of the main benefits is that section or line or even character segmentation can be completed within a single program itself. The contour function in MATLAB retrieves information regarding 'x' and 'y' coordinates of each contour. And with respect to the obtained 'y' values, further line segmentation is done.

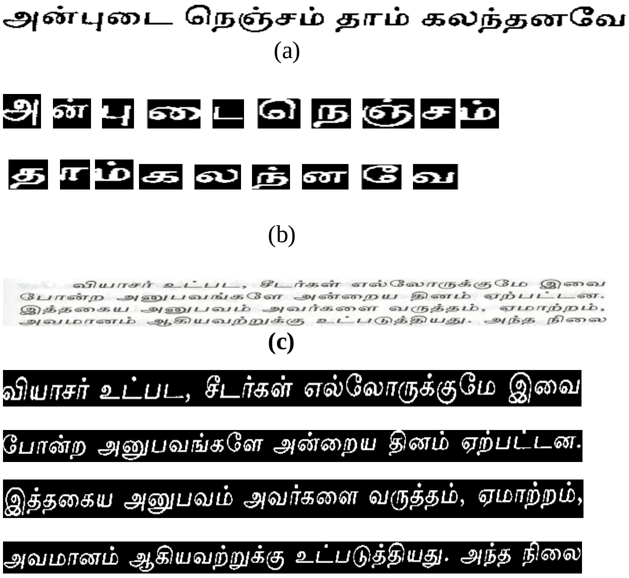
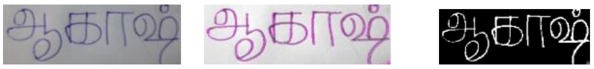

அன்புடை நெஞ்சம் தாம் கலந்தனவே
வியாசர் உட்பட, சீடர்கள் எல்லோருக்குமே இவை
போன்ற அனுபவங்களே அன்றைய தினம் ஏற்பட்டன.
இத்தகைய அனுபவம் அவர்களை வருத்தம், ஏமாற்றம்,
அவமானம் ஆகியவற்றுக்கு உட்படுத்தியது. அந்த நிலை
(a)
(b)
(c)
(d)

Fig 9: Illustration of Character & Line segmentation by Bresson's level set methodology (a) A sentence taken from a Tamil classic (b) Character segmented sample output for Bresson's method (c) A short paragraph from a Tamil novel (d) Line segmented output for Bresson's method.

4.5 Experiment with scanned hand-written Document



(a) (b) (c)

Fig 10: (a) Handwritten document (b) Segmentation result (c) Binarization result

Figure 10(a) contains a scanned handwritten document and Figure 10(b) and 10(c) shows the segmented and binarized output by Bresson's method. It is observed in majority of the cases that segmented output from the Bresson's algorithm is better than the segmented result obtained from Zhang's method. The segmented document of the Bresson's method obviously has more clarity than that of the Zhang's algorithm. Another reason for getting a clear output from the Bresson's algorithm is that it solves for a convex kind of minimization problem. This globally convex segmentation (GCS) method is both easier to handle numerically, and is more reliable because it does not stuck at local minima.

A total of 4 tables are illustrated to indicate the accuracy obtained for the binarization process by Bresson's method in case of Novel, Corrupted and hand written scanned documents. The following notations are used throughout the work:

NA - Novel Article

HA - Handwritten Article

CA - Corrupted Article

CR - Characters Retrieved

TC - Total Count

% - Accuracy in Percentage

LR - Lines Retrieved

LPP - Lack of Pre-Processing

EPP - Effect of Pre-Processing

Table 1. Binarization of novel articles by 'Bresson'

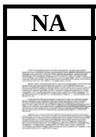
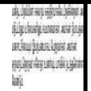
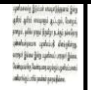
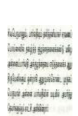
NA	CR	TC	%	LR	TC	%
	1101	1228	89.6	32	32	100
	150	163	92	5	5	100
	197	206	94.7	7	7	100
	157	164	95.7	6	6	100
	365	369	98.9	15	15	100

Table 1 shows the precision obtained by Bresson's method for good documents in Tamil. In case of neat documents, good accuracy is obtained without the need for any pre-processing phase. Line segmentation is done properly but the minimum accuracy of character segmentation for novel document is 89.6% and the maximum accuracy is high about 98.9%. Table 2 depicts the precision of binarization process in the absence and presence of Pre-processing phase comprising Coherence edge diffusion. Hand writing varies according to each person and hence we have tried to increase the results with diffusion. This is achieved in many cases, following with a significant improvement in the initial accuracy. Without diffusion, the highest precision obtained is 80%. By diffusion, the result for handwritten documents is elevated to 93.33%.Table 3 proves the substantial outcome of diffusion on binarization process in case of corrupted or noisy documents. This is evident in the last case of Table 3, where the document is noisy with 'Gaussian' of noise intensity level up to 0.4. In the absence of diffusion, binarization accuracy is around 45%. But once the noisy document is diffused, and then passed for binarization, the accuracy is attained at 74%. The average accuracy is obtained for different type of documents and is illustrated in Table 4. It's found from the experiments that novel documents get binarized well without the need for any diffusion. But in the case of corrupted and handwritten documents which may be noisy or degraded, the effect of pre-processing significantly improves the precision result of binarization procedure.

Table 2. Binarization of handwritten articles by Bresson's method in the absence and presence of 'pre-processing'

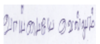
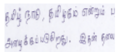
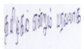
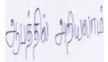
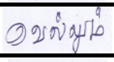
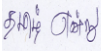
HA	TC	LPP		EPP	
		CR	%	CR	%
	12	6	50	9	75
	43	25	58.13	29	67.43
	15	12	80	14	93.33
	11	6	54.54	8	72.72
	5	2	40	3	60
	6	4	66.66	5	83.33

Table 3. Effect of Pre-processing on Binarization for Corrupted articles.

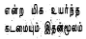
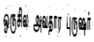
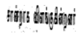
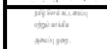
CA	TC	LPP		EPP	
		CR	%	CR	%
	19	9	47.36	12	63.15
	22	15	68.1	17	77.27
	13	10	76.92	11	84.61
	15	12	80	14	93.33
	31	14	45.16	23	74.19

Table 4. Average Accuracy for variety Tamil documents

Input Documents	Average Accuracy (%)	
	LPP	EPP
NA	94.18	-
CA	63.5	78.51
HA	58.22	75.30

5. CONCLUSION

This paper presents two powerful algorithm, Bresson's method and SBGFRLS method by Zhang, for binarization of documents in Tamil based on Level set technique. It is established that Bresson's algorithm works well for document binarization rather than Zhang's method. Another promising result obtained is that if the documents are old or degraded, or even hand written one, by the application of a Pre-processing phase, the quality of the binarized image is again upgraded. Here, in this paper, we have demonstrated the use of Coherence Edge diffusion as an intermediary step in case of corrupted documents so as to increase the accuracy of binarization results by Bresson's algorithm. Moreover, the accuracy obtained is enhanced by using methods like Total variation, Edge enhancement diffusion as a Pre-processing step.

6. REFERENCES

[1] N. Otsu, "A threshold selection method from Gray level histograms", *IEEE Transactions .on Systems, Man and Cybernetics*, Vol. SMC-9, No. 1, January 1979.

[2] Kaihua Zhang a, Lei Zhang a, Huihui Song, and Wengang Zhou, "Active contours with selective local or global segmentation: A new formulation and level set method", *Journal. Image and vision computing*, vol. 28, 2010.

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

[4] Maythapolnun Athimethphat, "A Review on Global Binarization Algorithms for Degraded Document Images", *AU J.T.* 14(3), January 2011.

[5] Bolan Su , Shijian Lu, and Chew Lim Tan, "Binarization of Historical Document Images Using the Local Maximum and Minimum", *ACM*, June 2010.

[6] Farjana Yeasmin Omee, Shiam Shabbir Himel and Md. Abu Naser Bikas, "A Complete Workflow for Development of Bangla OCR", *IJCA*, Volume 21– No.9, May 2011.

[7] Seethalakshmi R, Sreeranjini T R and Balachandar T, "Optical Character Recognition for printed Tamil text using Unicode", *J Zhejiang Univ SCI*, 2005.

[8] Scott Leishman, "Shape-Free Statistical Information in Optical Character Recognition", MSC Research Thesis, University of Toronto, 2007.

[9] Salem Saleh Al-amri, N.V. Kalyankar and N.V. Kalyanka "Image segmentation by using Edge Detection", IJCSE, Vol. 02, No. 03, 2010.

[10] Tom Goldstein, Xavier Bresson, and Stanley Osher, "Geometric Applications of the Split Bregman Method: Segmentation and Surface reconstruction", December 2009.

[11] Lisa Jonasson, Xavier Bresson, Patric Hagmann, Olivier Cuisenaire, Reto Meuli, and Jean-Philippe Thira, " White matter fiber tract segmentation in dt-mri using geometric flows*", Medical Image Analysis*, 9(9):223-236, 2005.

[12] 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, IEEE Computer Society*, page 244, Washington, DC, USA, 1996.

[13] A. Yezzi, S. Kichenassamy, A. Kumar, P. Olver, A. Tannenbaum, "A geometric snake model for segmentation of medical imagery", *IEEE Trans. on Med. Imag.*, 16(2):199-209, 1997.

[14] M. Kass, A. Witkin, and D. Terzopoulos. "Snakes: Active contours models*", International Journal of Computer Vision*, pages 321–331,1988.

[15] D. Mumford and J. Shah. "Boundary detection by minimizing functionals". *In Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 1985.

[16] S.K. Weeratunga, C. Kamath, "An Investigation of Implicit Active Contours for Scientific Image Segmentation", *Visual Communications and Image Processing Conference*, January 2004.

[17] Optical Character Recognition - Wikipedia, the free encyclopedia, http ://wikipedia:org/wiki/Tamil_alphabet"

[18] Luminita Vese, "An Introduction to Mathematical Image Processing", Under graduate summer school 2010, IAS, Park City Mathematics Institute, Utah.

[19] Haixia Wang, Qian Kemao, Wenjing Gao, Feng Lin, Hock Soon Seah, " Partial Differential Equation Based Coherence Enhancing Denoising for Fringe Patterns", *International Conference on Experimental Mechanics 2008*, Vol. 7375, 2008.

[20] Weickert J., "Coherence-Enhancing Diffusion Filtering", *International Journal of Computer Vision"*, vol. 31, issue 2-3, pp.111 - 127 (1999).

[21] Weickert J., V. Hlavac and R. Sara, Eds, "Multiscale Texture Enhancement", *Computer analysis of images and patterns*, Lecture Notes in Comp.Science, 970, Springer, Berlin, pp. 230-236(1995).

[22] H C Sateesh Kumar, K B Raja, Venugopal K R, and L M Patnaik, "Automatic Image Segmentation using Wavelets", *IJCSNS International Journal of Computer Science and Network Security*, Vol.9 No.2, February 2009.