Abstract

There are lots of ancient documents which are badly degraded. It is very difficult to segment text from the document, as there is a variation between the document background and foreground. Binarization technique that addresses these issues using adaptive image contrast. The adaptive contrast map is first constructed for degrade document. The contrast map is then binarized and combined with canny’s edge map to identify text stroke edge pixel.

1 Introduction

Nowadays, in the field of computer vision and information retrieval major topic is an image search. Image collection is increasing rapidly as enhance in image capturing devices like image scanners, digital cameras, mobile phone camera and also with increased use of multimedia data on internet. To deal with such huge image collection effective and efficient browsing, searching and retrieval tools are needed for different areas. In early era, various image retrieval systems had developed as: text-based and content-based. All text based image retrieval systems require the text description with images in large scale data bases and manually this task is not feasible. As a result, text based image retrieval systems were not applicable for task dependent queries [1]. To overcome these problems content based image retrieval (CBIR) was introduced in the early 1980s. Content based image retrieval (CBIR) is a technique used to retrieve the image based on features like color, texture and shape from large scale image dataset. In CBIR, images are indexed by their visual contents such as color, texture and shapes. Image retrieval process can be of two types- 1) Content based IR as low level features and 2) Semantic based IR as high level features. The tricky feature of CBIR is used to minimize the difference of content based features and semantic based features [2]. There exist lots of handwritten and printed historical manuscripts in libraries and museums in the world including medieval manuscripts, books of the renaissance, author manuscripts, old newspapers, archives, etc. [1]. Now day’s old and historical documents are archived and preserved in...
large quantities worldwide in digital form. A lot of research effort has been dedicated to optical character recognition (OCR) systems to convert paper information into digital. Numbers of algorithms are available for this purpose and many commercial OCR systems are now available in the market. But most of these systems work on good quality documents. There are not sufficient numbers of research works for OCR that can handle degraded or poor quality handwritten and printed documents. Pre-processing is the most important steps of OCR systems. Pre-processing of poor quality documents is very important for the accuracy of the other phases of OCR. Binarization (thresholding) of document images is the first most important step in pre-processing of poor quality scanned documents to save all or maximum subcomponents such us text, background and image [2]. Binarization computes the threshold value that differentiate object and background pixels [3]. The main advantage of binary images is that it decreases computational load and increases efficiency of the systems. Binary images can be obtained from gray level images by binarization. Binarization of degraded document images is not an easy task. The binarization is become more difficult when there are a varying illumination, variance of gray levels within the background and the object, inadequate contrast, object shape, noise and size non-commensurate with the scene. Degradations appear frequently and may occur due to several reasons which range from the acquisition source type to environmental conditions.

2 Edge Detection Techniques

A. Edge Detection:- Canny finds edges by looking for local maxima of the gradient of f(x, y). The gradient is calculated using the derivative of a Gaussian filter. The method uses two thresholds to detect strong and weak edges, and includes the weak edges in the output only if they are connected to strong edges. Therefore, this method is more likely to detect true weak edges. The implemented Canny edge detector presented the best performance. Both visually and quantitatively based on the measures such as mean square distance, error edge map and signal to noise ratio. Using the implemented canny edge detector as an enhancement tool for remote sensing images, the result was robust and achieved a very high enhancement level [3].

B. Complement Image:- This process is to transform an edge image into regions as well as reduce unnecessary noise such as spurs to provide a clean region image. Applying binary invert operator to the binary edge image gives a region image in binary with 1’s where a region found a 0’s elsewhere is given. In the complement of a binary image, black becomes white and white becomes black. This operation facilitates further morphological filtering and smoothing operations which are carried out in next phases for fine edge detection.

C. Morphological Open operator:- Morphology relates to structure or form of objects. Morphological filtering simplified segmented images by smoothing out object outlines filling small holes, eliminating small projections. Primary operations are dilation and erosion. These operations use a structuring element which determines exactly how object will be dilated or eroded. Dilation process expands image objects by changing pixels with value of "0" to "1". On the other hand the erosion process shrinks binary objects by changing pixels with a value of "1" to "0". There is also a combination of dilation and erosion called opening and closing. Opening is erosion followed by dilation.

D. Morphological Erosion:- The morphological erosion operator is applied for localisation of edges image that shrinks region pixels in size which make clear image. Figure shows the comparison of the edge detections for the example image.
3 Review of Binarization Technique

Ioannis P 1 et al. (2013) [8] has discussed a contest. The general target of the contest is to distinguish current developments in document image binarization for both machine-printed and manually written record images utilizing assessment execution measures that obey document image investigation and recognition. The contest details have been described incorporating the assessment measures used and also the execution of the 23 submitted routines as well as a short depiction of every strategy.

A Sehad et al.(2013) [9] has present a capable scheme for binarization of ancient and degraded document images, grounded on texture qualities. The suggested technique is an adaptive threshold-based. It has been calculated by using a descriptor centred on a co-occurrence matrix and the scheme is verified objectively, on DIBCO dataset degraded documents furthermore subjectively, utilizing a set of ancient degraded documents offered by a national library. The outcomes are acceptable and assuring, present an improvement to classical approaches.

H Z Nafchi et al.(2013) [10] has concluded that the pre-processing and post processing phases meaningfully advance the performance of binarization approaches, particularly in the situation of harshly degraded ancient documents. An unverified post processing technique is presented founded on the phase-preserved denoised image and also phase congruency features extracted from the input image. The central part of the technique comprises of two robust mask images that can be used to cross the false positive pixels on the production of the binarization technique. Firstly, a mask with an extreme recall value is attained from the denoised image with the help of morphological procedures. In parallel, a second cover is acquired dependent upon stage congruency features. At that point, a median filter is utilized to evacuate noise on these two masks, which then are utilized to rectify the yield of any binarization strategy.

J Parker et al.(2013)[11] has studied that regularly documents of notable noteworthiness are ran across in a state of deterioration. Such archives are regularly examined to all the while history and announce a disclosure. Changing over the data found inside such reports to open information happens all the more rapidly and inexpensively if a programmed technique to upgrade these corrupted archives is utilized as opposed to improving each one document image by hand. A novel mechanized image upgrade approach that indulges no preparation information was introduced. The methodology was valid to images of typewritten text in addition to hand written text or both.

Konstantinos N et al.(2013) [12] has analysed that document image binarization is of incredible value in the document image examination and recognition pipeline as it disturbs further phases of the recognition procedure. The assessment of a binarization technique helps in examining its algorithmic conduct, and also confirming its adequacy, by giving qualitative and quantitative sign of its execution. A pixel-based binarization assessment approach for recorded handwritten/machine-printed document image has been proposed. In the proposed assessment procedure, the review and accuracy assessment measures are fittingly adjusted utilizing a weighting plan that decreases any potential assessment unfairness. Extra execution measurements of the proposed assessment plan comprise of the rate rates of broken and missed content, false alerts, foundation commotion, character amplification, and combining.
V Rabeux et al. (2013) [13] has an approach to expect the outcome of binarization algorithms on a known document image according to its situation of degradation. Document shaving degradation which result in binarization errors. To characterize the degradation of a document image by using different features based on the strength, amount and position of the degradation. These characteristics allow us to build calculation models of binarization algorithms that are very accurate according to $R^2$ values and $p$-values. The prediction models are used to select the best binarization algorithm for a given document image.

D GACEB et al. (2013) [14] has studied a smart binarization technique of the images. In this technique, considered different degradations document images. The nature of every pixel is approximate using a hierarchical local thresholding in order to classify it as foreground, background or ambiguous pixel. The ambiguous pixels that represent the corrupted zones cannot be binarized with the same local thresholding. The global quality of the image is estimated from the density of theses degraded pixels. If image is degraded then apply a second separation on the ambiguous pixels to split them into background or foreground. Second process uses our improved relaxation method.

Marian W et al. (2013) [15] has implemented a quick and proficient document image clean up and binarization technique depend on retinex hypothesis and global thresholding. This technique joins of local and global thresholding with concept of retinex theory which can efficiently improve the degraded and poor quality document image. Then, quick global threshold is utilized to change over the document image into binary form. The new method conquers the limitations of the related global threshold techniques.

Vassilis P et al. (2012) [16] has discussed an capable technique dependent upon mathematical morphology for extracting text regions from degraded document images. The fundamental stages of methodology area) top-hat-by-reconstruction to construct a filtered image with sensible background) region growing beginning from a set of seed points and attaching to each seed similar intensity neighbour pixels and c) conditional extension of the first detected text regions based on the values of the second derivative of the filtered image.

Bolan Su et al. (2012) [17] has studied a document image binarization structure that makes utilization of the Markov Random Field model. Structure isolates the document image pixels into three classes i.e. document background text, document foreground text, and unknown pixels established binarization method. Unknown pixels are belonging to foreground and background categories by incorporating MRF model and boundary information.

C. Patvardhan et al. (2012) [18] has studied that images may contain difficult background i.e. shading or a denoising. Binarization method of document images creates them suitable for OCR using discrete curve let transform. Curve let transform is used for eliminate difficult image background, white Gaussian noise and gives improved binarized image. The Curve let transform also helps to enhanced in text shape still in the occurrence of noise. This method is capable to eliminate high frequency Gaussian noise and low frequency complex backgrounds and shows better performance.
4 Drawbacks
Various techniques have been put forth so far for document binarization as shown in literature survey. It has been concluded from the existing research is that no technique is perfect for every case. Therefore still research is required in this field of image binarization. Following are the main limitations

➢ Filters were used by many researchers to reduce the noise from the image but the use of the Decision based switching filter (best edge preserving filter) is not found. It may increase the accuracy of the available binarization methods
➢ Contrast enhancement is either done by tradition methods or not done. So adaptive contrast enhancement is required.
➢ Most of the methods have neglected the use of edge map which has the ability to map the exact character in efficient manner.

5 Conclusion
This paper has focused on the ancient degraded document binarization technique. The main objective of this paper is to evaluating the short comings of algorithms for degraded image binarization. It’s found that each technique has its own benefits and limitations; no technique is best for every case. The main limitation of existing work is found to be noisy and low intensity images. In near future we will propose a new algorithm which will use more reliable methodology to enhance the work.

6 References
Bency Jacob, Prof. S. B. Waykar: A Survey On Binarization Of Historical Degraded Documents

Paper ID: IJIFR/ V2/ E2/ 026


