Locating the Query Block in a Source Document Image

Naveena M and G Hemanth Kumar
Department of Studies in Computer Science, University of Mysore, Manasagangotri-570006, Mysore, INDIA.

Abstract: **- In automatic document analysis is the discrimination text images. This is for the segmentation of text images in digitized documents. In this method mainly working based on the representation of window-like portions of a document by means of their gray level histograms. Through empirical evidence it is shown that text images regions have different gray level histograms. Unlike the usual approach for the characterization of histograms that is based on statistics parameters. This approach works with the histogram normalization, cumulative histogram, and Euclidian formula. since it possesses all the information contained in the histogram pattern. The next and logical step is to automatically select the most discriminant spectral components as far as the text images segmentation goal is concerned. A fully automated procedure for the optimal selection of the discriminant features is also expounded.**

Keywords**: Scanned and printed document images, Image Compression, Edge, and Classification.**

I. INTRODUCTION

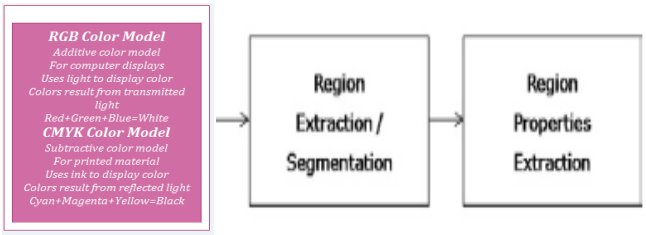
Content-based image retrieval (**CBIR**), also known as query by image content (QBIC) and content-based visual information retrieval (CBVIR) is the application of computer vision to the image retrieval problem, that is, the problem of searching for digital images in large databases.

"Content-based" means that the search will analyze the actual contents of the image. The term 'content' in this context might refer colors, shapes, textures, or any other information that can be derived form the image itself. Without the ability to examine image content, searches must rely on metadata such as captions or keywords. Such metadata must be generated by a human and stored alongside each image in the database.

It is the application of computer vision techniques to the image retrieval problem, specifically the search for specific digital images in large databases The two approaches commonly used for image retrieval are referred to simply as global-based image searches and region (or sub-image)-based image searches. An important distinction between these approaches is that global-based methods enable whole image matching and consider how much of an image is relevant, while region-based methods focus primarily on specifying a region and on retrieving a large number of images with similar objects. Both methods are useful for image retrieval, but are best suited to queries of different types. Searching by global distinction is the preferred approach in cases where the user provides a whole image for query, where queries take the form of ‘‘show me more relevant images that look like this query image’’.

However, if the user is interested in finding something located in a specific part of an image (e.g., ‘‘show me relevant images with a red flower on the right’’), global-based retrieval is unable to resolve spatially localized color regions from the global distribution and region based image searches will be more successful. For both these techniques, the retrieval system must incorporate a function capable of performing the automated extraction and efficient representation of visual features.

There are two different kinds of tasks involved in this process: object-presence detection and object localization .Object-presence detection seeks to determine whether one or more objects are present anywhere in the image.



Extraction of this information involves detection, localization, tracking, extraction, enhancement, and recognition of the text from a given image.

Content-based image indexing refers to the process of attaching labels to images based on their content. Image content can be divided into two main categories: perceptual content and semantic content

II. PROPOSED ARCHITECTURE

The machine printed text and scanned image is considered to spot the words. Scanned image may contain noise; to remove the noises is a challenging task. Then word segmentation is carried out to calculate features for each word to spot the desires word. For this, a Model is proposed. The proposed architecture is simple to use and understands. The architecture is as shown in the following block diagram. Each model; in the architecture is explained in the section 2.2.

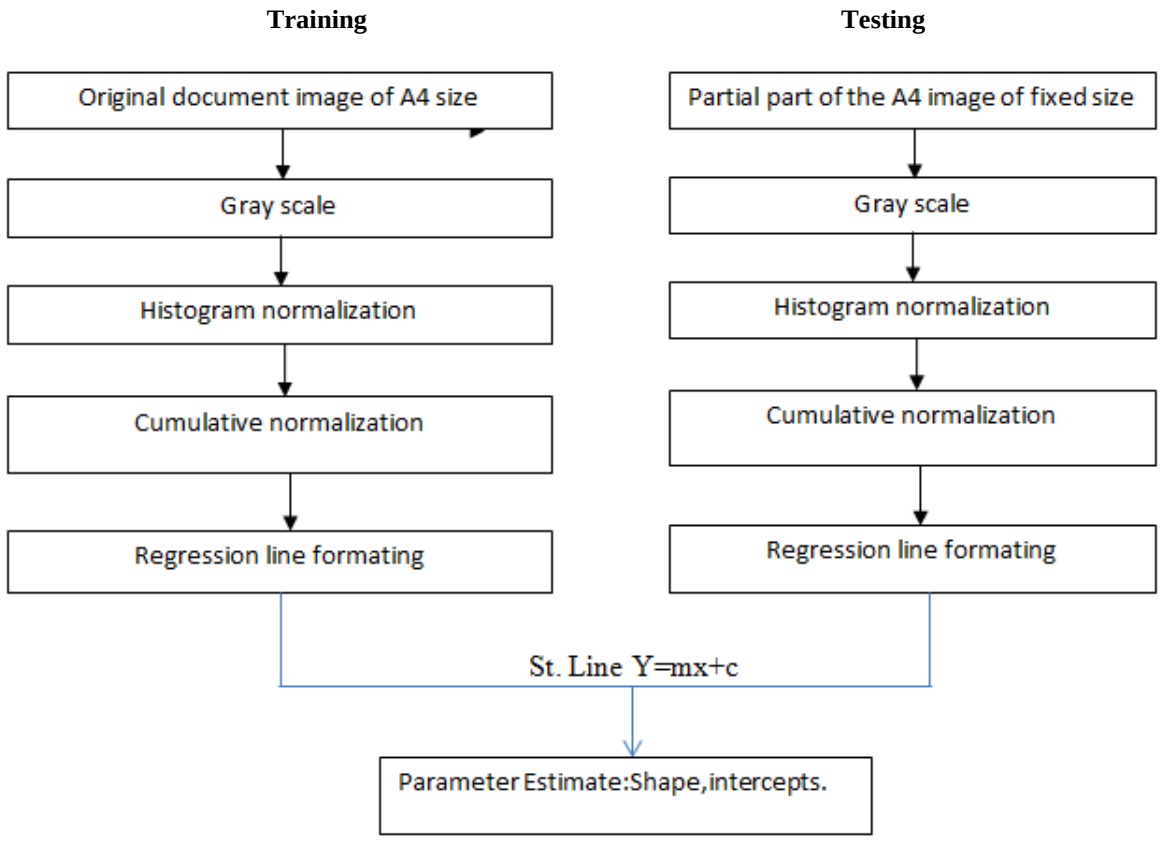


Fig 2.1: Structural view of Proposed System

2.1 Design Issues

2.1.1 Pre-processing:

Digital images are prone to a variety of types of noise. Noise is the result of errors in the image acquisition process that result in pixel values that do not reflect the true intensities of the real scene. There are several ways that noise can be introduced into an image, depending on how the image is created. For example:

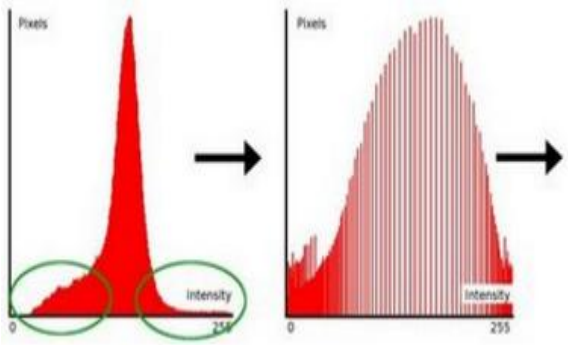
- If the image is scanned from a photograph made on film, the film grain is a source of noise. Noise can also be the result of damage to the film, or be introduced by the scanner itself.
- If the image is acquired directly in a digital format, the mechanism for gathering the data (such as a CCD detector) can introduce noise.
- Electronic transmission of image data can introduce noise.

Scanned document image is taken to spot the words. The scanned document image is binarized for further processing. The scanned image may contain some noises. By using Median Filtering, noises in the scanned document image can be removed.

Median filtering is a nonlinear operation often used in image processing to reduce noise. Median filtering is similar to that of an averaging filter, in that each output pixel is set to an average of the pixel values in the neighborhood of the corresponding input pixel. However, with median filtering, the value of an output pixel is determined by the *median* of the neighborhood pixels, rather than the mean. The median is much less sensitive than the mean to extreme values (called *outliers*). Median filtering is therefore a better way to remove these outliers without reducing the sharpness of the image. A median filter is more effective than convolution when the goal is to simultaneously reduce noise and preserve edges. Then dilating the noise removed image fixing structure element. The dilation is one of the operations in mathematical morphology. The dilating operation usually uses a structuring element for probing and expanding the shapes contained in the input image.

III. EXPERIMENTAL ANALYSIS

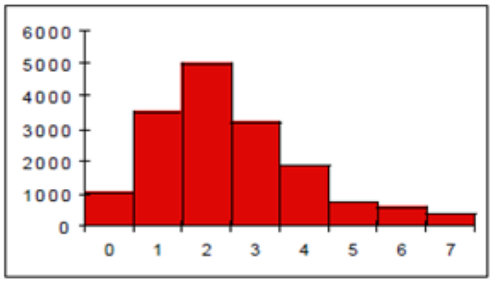
Consider a 128*128 pixels image that contains L=8 gray levels with the following distribution of pixels.



RGB Color Model
Additive color model
For computer displays
Uses light to display color
Colors result from transmitted light
Red+Green+Blue=White

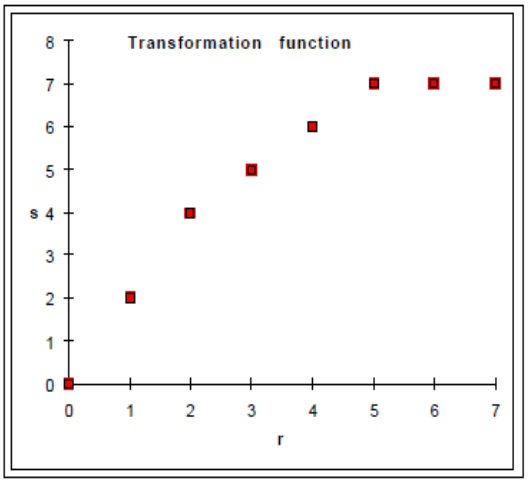
CMYK Color Model
Subtractive color model
For printed material
Uses ink to display color
Colors result from reflected light
Cyan+Magenta+Yellow=Black

gray levels r_k	n_k
0	1028
1	3544
2	5023
3	3201
4	1867
5	734
6	604
7	383
	16384



gray levels r_k	n_k	n_k / N	$\sum_{j=0}^{k} n_j / N$	$(L-1)\sum_{j=0}^{k} n_j / N$	new gray levels s_k
0	1028	0.0627	0.0627	0.4392	0
1	3544	0.2163	0.2791	1.9534	2
2	5023	0.3066	0.5856	4.0994	4
3	3201	0.1954	0.7810	5.4670	5
4	1867	0.1140	0.8950	6.2647	6
5	734	0.0448	0.9398	6.5783	7
6	604	0.0369	0.9766	6.8364	7
7	383	0.0234	1.0000	7.0000	7
N=	16384	1			

The new equalized gray levels:		
T(0) =	0	
T(1) =	2	
T(2) =	4	
T(3) =	5	
T(4) =	6	
T(5) =	7	
T(6) =	7	
T(7) =	7	

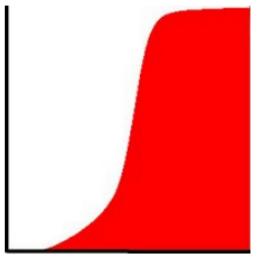


- Equalization implies *mapping* one distribution (the given histogram) to another distribution (a wider and more uniform distribution of intensity values) so the intensity values are speeded over the whole range.
- To accomplish the equalization effect, the remapping should be the *cumulative distribution function (cdf)*

  For the histogram $H(i)$, its *cumulative distribution* $H'(i)$ is:

$$H'(i) = \sum_{0 \le j < i} H(j)$$

To use this as a remapping function, we have to normalize $H'(i)$ such that the maximum value is 255 (or the maximum value for the intensity of the image). From the example above, the cumulative function is:



Finally, we use a simple remapping procedure to obtain the intensity values of the equalized image:

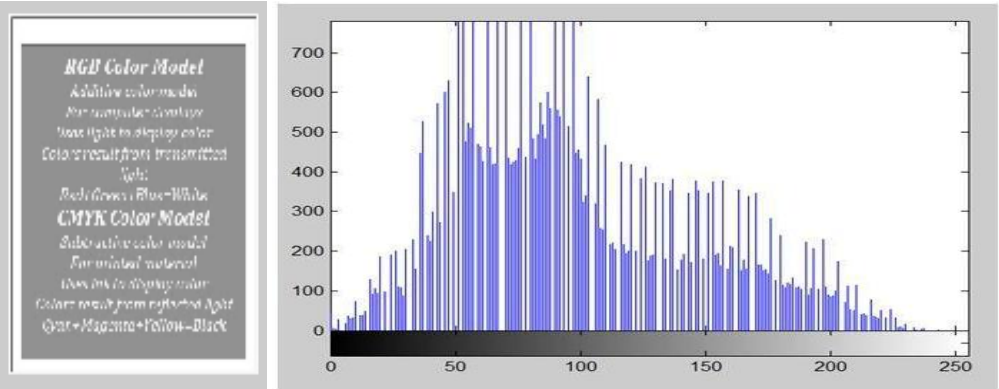
$$equalized(x,y) = H'(src(x,y))$$

Histogram equalization is used to enhance contrast. It is not necessary that contrast will always be increase in this. There may be some cases were histogram equalization can be worse. In that cases the contrast is decreased.

Lets start histogram equalization by taking this image below as a simple image.

The histogram of this image has been shown below.

Image



3.1 CDF:

Our next step involves calculation of CDF (cumulative distributive function). Again if you donot know how to calculate CDF , please visit our tutorial of CDF calculation.

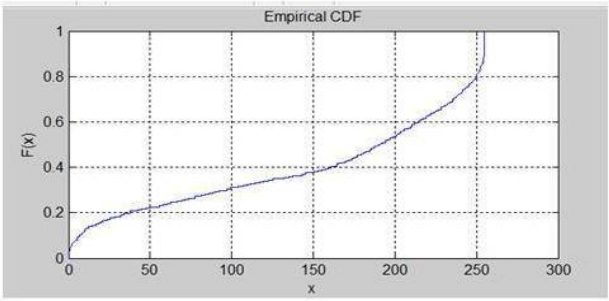
Lets for instance consider this , that the CDF calculated in the second step looks like this.

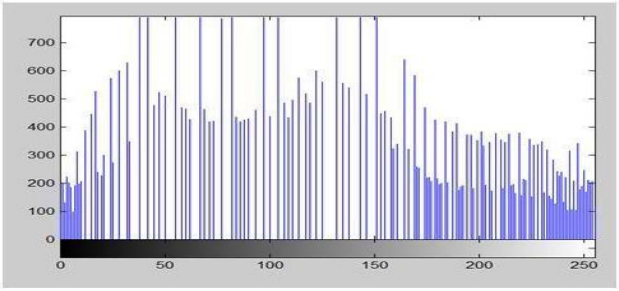
Gray Level Value	CDF
0	0.11
1	0.22
2	0.55
3	0.66
4	0.77
5	0.88
6	0.99
7	1

Lets assume our old gray levels values has these number of pixels.

Gray Level Value	Frequency
0	2
1	4
2	6
3	8
4	10
5	12
6	14
7	16

CUMULATIVE DISTRIBUTIVE FUNCTION OF THIS IMAGE





HISTOGRAM EQUALIZATION HISTOGRAM

3.2 Methodology

An intensity histogram is a graph, plotting the number with a specific gray level vs. the gray level value. Normalize an histogram is a technique consisting into transforming the discrete distribution of intensities into a discrete distribution of probabilities.

$$n_{kn} = \frac{n_k}{length \times width} = p_r(r_k)$$

Which could be written in terms of mathematical transformation:

$$\begin{cases} [0, L-1] \to \mathbb{N} \\ x \to Card(x) \end{cases} \quad \text{becomes} \quad \begin{cases} [0, L-1] \to [0,1] \\ x \to pdf(x) = \frac{Card(x)}{\sum_{i=0}^{L-1} Card(x_i)} \end{cases}$$

Where *Card* means the cardinality of the set so in our case the number of pixel.

$$m = r\left(\frac{s_y}{s_x}\right)$$

IV. CONCLUSION AND FUTURE WORK

One can use scanned image queries to retrieve math expressions from document databases using page segmentation and image similarity algorithms, by which optical character recognition can be avoided.

Today information technology has proved that there is a need to store, query, search and retrieve large amount of electronic information efficiently and accurately. So document image retrieval is very challenging field of research with the continuous growth of interest and increasing security requirements for the development of the modern society. This paper surveys the technical achievements in the field of document image retrieval, discuses system architecture, comprehensive survey of various proposed methods to retrieve the documents. It also highlights the challenges and scope of research.

REFERENCES

[1]. R. Datta, D. Joshi, J. Li, and J. Z. Wang, "Image Retrieval: Ideas, Influences, and Trends of the New Age," ACM Computing Surveys 40, 2 (2008).
[2]. N. Vasconcelos, "From Pixels to Semantic Spaces: Advances in Content-Based Image Retrieval," Computer 20, 20–26 (2007).
[3]. J. Ha, R. M. Haralick, and I. T. Phillips, "Recursive x-y cut using bounding boxes of connected components," Proceedings of the Third International Conference on Document Analysis and Recognition 2, 952 (1995).
[4]. G. Nagy and S. Seth, "Hierarchical representation of optically scanned documents," Proc. of ICPR pp. 347–349 (1984).
[5]. T. Rath and R. Manmatha, "Word image matching using dynamic time warping," IEEE Computer Society Conference on Computer Vision and Pattern Recognition 2, 18–20 (2003).
[6]. .Likforman-Sulem, L., Zahour, A. and Taconet, B.,"Text line Segmentation of Historical Documents: aSurvey", International Journal on Document Analysis and Recognition, Springer, Vol. 9, Issue 2, pp.123-138, 2007.
[7]. U. Pal and P. P. Roy, "Multi-oriented and curved textlines extraction from Indian documents", IEEETrans. On Systems, Man and Cybernetics- Part B,vol. 34, pp.1676-1684, 2004.
[8]. U. Pal, B.B. Chaudhuri. (2004): Indian script character recognition: a survey, Pattern Recognition,37, 1887 – 1899.
[9]. B. B. Chaudhuri and U. Pal, "A complete printed Bangla OCR system", Pattern Recognition, vol.31, pp.531-549,1998.
[10]. K. Wong, R. Casey and F. Wahl "Document Analysis System ", IBM j.Res .Dev., 26(6), pp.647-656, 1982.
[11]. Grundland M, Dodgson N, (2007) Decolorize: Fast, contrast enhancing, color to grayscale conversion. Pattern Recognition 40: 2891–2896.
[12]. Cadik M, (2008) Perceptual evaluation of color-to-grayscale image conversions. Computer GraphicsForum 27: 1745–1754.
[13]. Rafael C. Gonzalez, Richard E. Woods, (2007) "Digital Image Processing", 2nd ed., Beijing:Publishing House of Electronics Industry.
[14]. Zimmerman, JB, SM Pizer, EV Staab, JR Perry, W McCartney, BC Brenton, (1988) "An Evaluationof the Effectiveness of Adaptive Histogram Equalization for Contrast Enhancement", IEEE Trans.Med. Imaging, 7(4): 304-312.
[15]. N. Ezaki, M. Bulacu, L. Schomaker, (2004) "Text Detection from Natural Scene Images: Towards aSystem for Visually Impaired Persons", Int.Conf. on Pattern Recognition (ICPR), vol. II, pp. 683-686.
[16]. J. Park, G. Lee, E. Kim, J. Lim, S. Kim, H. Yang, M. Lee, S. Hwang, (2010) "Automatic detectionand recognition of Korean text in outdoor signboard images", Pattern Recognition Letters.