Journal of Nonlinear Analysis and Optimization
Vol. 11(10) (2020), October 2020
https://ph03.tci-thaijjo.org/
ISSN : 1906-9685

J.Nonlinear
Anal. Optim

An Overview of Document Image analysis

Dr.T.Krishnamoorthy
Associate Professor, Department of ECE
Sri Sai Institute of Technology and Science, Rayachoti
Email: krishnamoorthyphd@gmail.com

Abstract :Modern Technology has made it possible to produce,process, transmit and store digital images efficiently.Consequently the amount of visual information is increasing at an accelerating rate in many diverse application areas.The large amount of these image data are related to text.The information is stored in the form of digital versions and in document management system .Document image retrieval systems are utilized in many organizations which are using Document image databases extensity.To fully exploit this new content based image retrieval techniques are required.

Introduction :When we refer to a paper document it is distinguished by the fact that it is on paper.However the notion of digital document is one which digital systems can understand and present to the user in an articulated manner. There are several types of documents which present information to a person that can be conceived and comprehended. Documents can be primarily divided into three different categories and they are
Online: Documents that fall into the online paradigm consist of online handwriting that consist handwriting data captured by a digitizer that captures handwriting of a writer.These digitizers are specialized devices that capture a writer's ink information his speed the pressure applied etc.which can be later used for further processing

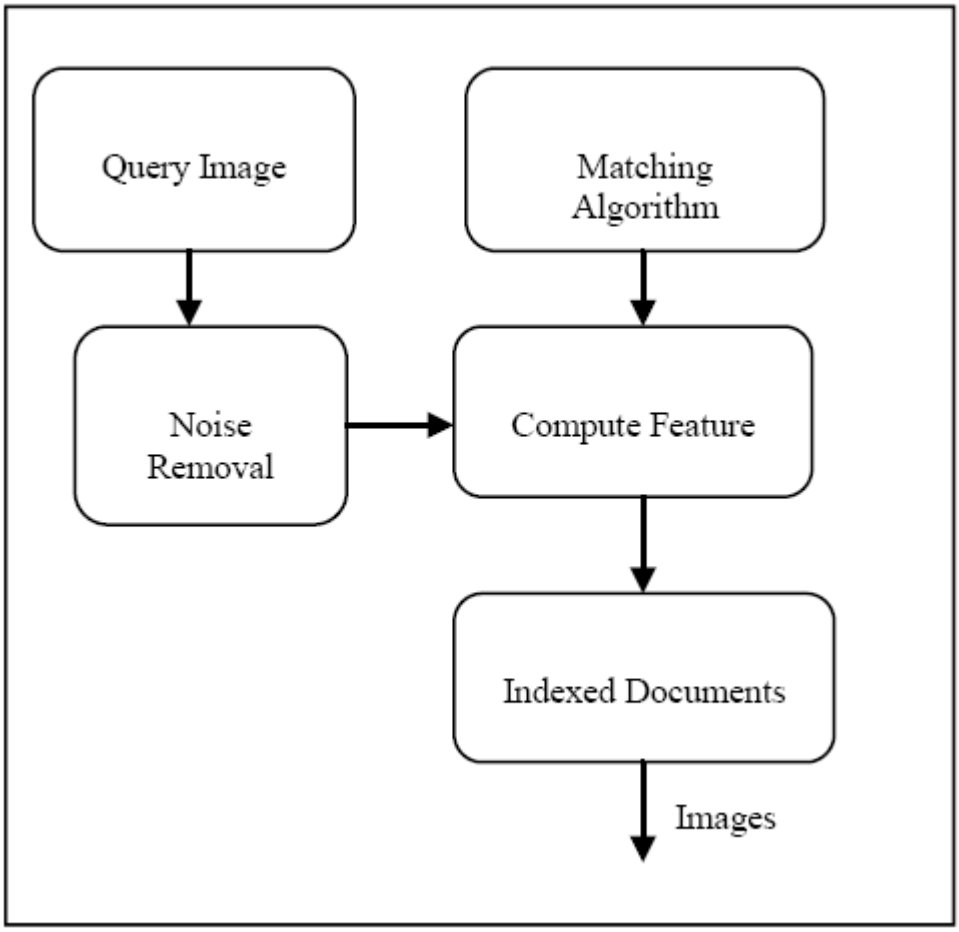
Offline: The least common denominator for handwriting is the paper and pen. Offline documents consist of scanned

copies of handwriting information that were a priori written on a sheet of paper.Scanner assumes the role of digitizer and scan the document after the writer has written his content., we donot access to information such the speed or pressure with which the writer must have written.Handwriting data in both online or offline could have cursive,discrete or mixed.Handwritten letters personal notes are examples of offline documents.

Printed:Printed documents contain textual information which are scanned copies from a book.The textual content in books are in the printed form following a specific font style font size and maintaining a standard uniformity across all pages.These are typically referred to as document images and could also contain images or pictures in addition to textual content.OCRs are used to extract the textual content from these documents,Books,Journals,articles,newspapers,magazines are some of the examples of a

printed document.Some of the popular digitizers for both offline and printed documents include the digital cameras,hand held and flat bed scanners etc.

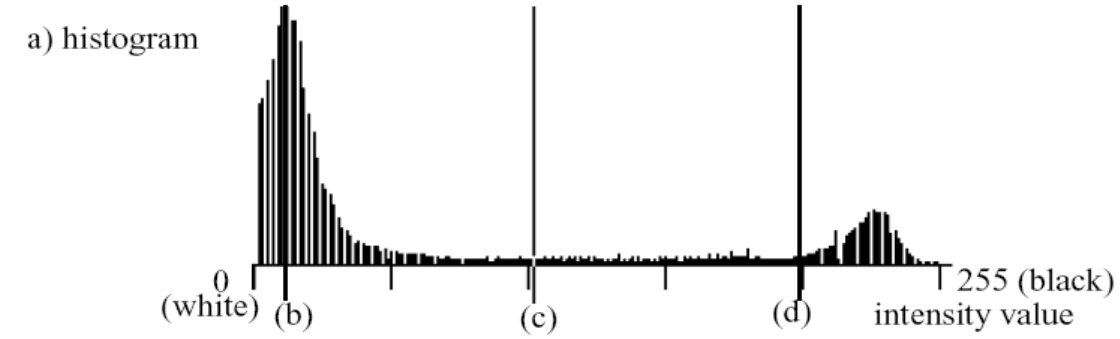
Document image retrieval: The various steps in document image retrieval system are feature extraction,Noise removal and matching algorithm



Document image retrieval system

Data capture of documents by optical scanning or by digital video yields a file of picture elements or pixels that is the raw input to document analysis.These pixels are samples of intensity values taken in a grid pattern over the document page where the intensity values may be :OFF(0) or ON(1) for binary images,0-255 for gray scale images and 3 channels of 0-255 color values for color images. The first step in document analysis is perform processing on this image to prepare it for further analysis.Such processing includes thresholding to reduce a gray scale or color image to binary image,reduction of noise to reduce extraneous data and thinning and region detection to enable easier subsequent detection of pertient features and objects of interest.

Query Image: Query Image is a request form end user for retrieval indexed documents . First end users enter query image, then system retrieval document images relevant with query image.



b) low threshold

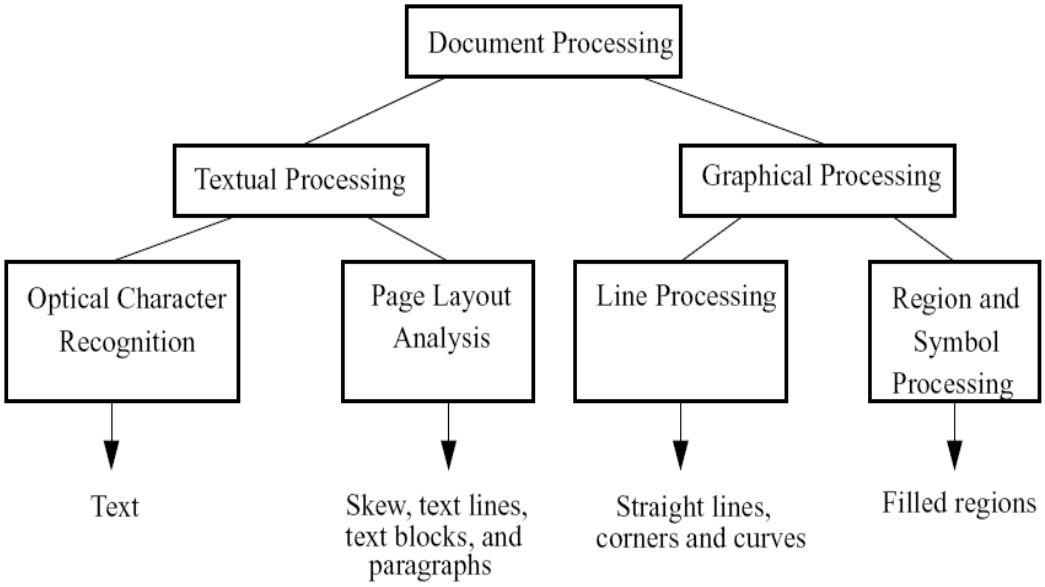
c) good threshold

d) high threshold

Image binarization. a) Histogram of original gray-scale image. Horizontal axis shows markings for threshold values of images below. The lower peak is for the white background pixels, and the upper peak is for the black foreground pixels. Image binarized with: (b) too low a threshold value, c) a good threshold value, and d) too high a threshold value.

Noise removal: Noise removal is fulfilled to get remove of each noise or printed text extending over the extracted images such as logos, ignature, achine print etc.In the preprocessing stage the printed text is dismissed from the image instances. To dismiss the printed text form images various methods can be used as example of image improvement methods based on SVM,Chain Code, to classify each connected component as a part of signature,Noise elements , logo, handwritten text,Noise etc.

Word segmentation stage follows the preprocessing stage.Its primary goal is to detect the word limits and filter the noise and punctuation marks.This is accomplished by using the connected components labeling and filtering method.Since the noise of some documents can change significantly the shape of the extracted word images, the noise and the punctuation points must be rejected.



A Hierarchy of Document image processing

Feature Extraction: Feature extraction includes extracting the significant knowledge from document images. One time the features are extracted they are saved in the data base. One of the biggest benefits of feature extraction is that meaningfully decreases the information to represent an image for comprehending the content of that image.It uses variety technical skills to extract the features like for example structural,concavity features and gradient which

criterions the image attributes at local,medium and large scale, features based on key block features and density distribution.Angular radial portioning of a images regions fisher classifier,DTW, conditional random field etc. are used for feature extraction.
Matching Algorithm: The document image retrieval is executed use of similarity method to compare the query with image database.
1 Noise removal from the query
2. Feature extraction from the query image.
3. Matching the query image features to each of the documents that indexed in the data base.
4. Sorting the documents in accord with the results from the Matching method.The work of matching algorithm is to contrast the feature with the features of the document images.
Similarity measure the data base feature vector and query feature vector is compared use of distance measure.The similarity of different metrics like for example chebychev,Euclidean, Manhattan etc. is done in. The normalized similarity is believed to be good for feature vectors as characterization to other measures.The Euclidean distance between the features of the query image and the indexed features in the databases involved in the document is computed.

$$Dis(p,r) = \sqrt{\sum_{i=1}^{n} (Q(p_i) - D(p_i, r))^2}$$

Where p is the feature that is being compared, D is the feature of the document.Q is the feature of the query, n is the count of component of the feature vector and r is the quantity of the document compared query.Eventually there is a set Dis(p.r) which comprise of the Euclidean distances between each indexed document and the query for any features .
Indexed Documents: Indexed documents are That display to user as results.

Evaluation Metrics: Document image retrieval is subset of information retrieval system.
Two most common and fundamental metrics for information retrieval impressiveness is precision and recall
Precision (P) is the count of retrieved documents that are relevant

$$Precision = \frac{\#(\text{relevant items retrieved})}{\#(\text{retrieved items})} = P(\text{relevant}|\text{retrieved})$$

Recall is the count of relevant documents that are retrieved

$$Recall = \frac{\#(\text{relevant items retrieved})}{\#(\text{relevant Items})} = P(\text{retrieved}|\text{relevant})$$

Evaluations of Document image retrieval : The measures that are considered for document image retrieval and indexing approaches are as follows.
Application Type: A document image retrieval approach has various applications such as similarity documents, word searching,duplicate detection etc.
Appearance Features: A document image retrieval approach has many appearances features. Any approach has specified appearance feature for it.
Query image : any approach has query image for retrieval documents such as signature image or word image.First users enter query image, then system retrieval document images relevant with query image.
Is Structural: Meaning of structural is table and formatting in document images.
Language Independent: The approach sutible for document image retrieval language independent.
Cost: Searching form large collection of document images passes through many steps: Image processing,feature extraction, matching and retrieval of documents. Each of these steps could be cost expensive.Each of the approach has different cost for matching and retrieval .

Techniques: Each of the approach has various techniques for indexing and retrieval documents.
Problems: Complex document pose a great challenge in the filed of document recognition and retrieval.Each of the approach has different problems such as noisy data,uncommon fonts etc.

Conclusion: Traditionally transmittal and storage of data have been by paper documents. In days gone by few ten years documents more and more begin on the computer but despite this it is vague whether the computers has enlarged or reduce the quantity of paper.Despite the fact that the concept of raw document image retrieval is interesting, inclusive resolutions which donot demand finish and exact conversion to a machine readable form continue to be evasive for feasible systems.Many approaches come in for indexing and retrieval document images. In this paper a frame work of document image analysis and its retrieval evaluation metrics are focused

References:

Edwards,J.Teh,Y.W.,Forsyth,D.Bock,R.,Maire,M.Vesom,G.Making Latin manuscripts searchable using gHMM's in:Proceedings of the 18th Annual Conference on Neural information processing Systems,Cambridge,USA,2004,pp.385-392.

Isthiani Y.Model-based information extraction method tolerant of OCR errors for document images In :Proceedings of the sixth international conference on Document Analysis and Recognition,Seattle,USA,2001,pp.908-915

Kavallieratou,E,Fakotakis,N,kokkinakis,G,2002.Un off-line unconstrained handwriting recognition system.International journal of Document Analysis and Recognition 4,226-242