OPTIMIZED SEARCH ENGINE TO FIND IMAGE BY PROVIDING KEYWORD

Prof.Gunjan Agre, Prof. Jagdish Pimple, Vaibhav Bhavsar, Vipeen Sarode, Palash Dhande
Department of Information Technology, Nagpur Institute of Technology
gunjan.agre@gmail.com
vaibbhavsar@gmail.com
vipeensarode@gmail.com
palash14@gmail.com

Abstract— **In this project we are developing a search engine to find the image by providing keyword the text is mined from image. The user enter the keyword for search, if that keyword is present in image which is available in database then the text is mined from image with the help of extraction technique and the searching and sorting algorithm is used to gives the result. Number of image are sorted according to the keyword is present in image.**

I. Introduction

In this project developing a technique where the text is mined from image. User can able to find the image by providing the text where require keyword is present. Text localization and recognition in images is important for searching information in image. To obtain satisfactory results, using the optical character recognition (OCR) technique which help to extract text from images. The Optical Character Recognition is the process of identification of texts from images or electronic document.

Basically user wants to access different kinds of data like text, images and documents. If user wants to access image by providing keyword, there is no particular way to do that.

To overcome this problem introducing methodology, where user can able to access all image which contain the searched keyword.

It will sort such images from the database which contain keyword on the images given by the user. To do this we use Optical Character Recognition technique, which extract texts from images.

The database and web both are large scale area or sources which contain the large amount of data, it is important to find methods to annotate and organize image databases in proper manner. The reason of developing perfect and robust image annotation model is presenting result as a bunch of images where the required keyword is present and fulfills the requirement of the user.

The most of the time when people search keyword, they got result in the form of text, documents. Here we are developing a technique which will extract the text from images. Extracting text from an image is not a difficult process, but it does require some tinkering with the problem at hand. There are a few procedures you can use. The basic procedure is to identify "text" versus background" pixels then extract individual symbol.

The complication comes when, the image quality is poor or the pixel cannot be well distinguished from the background. To avoid this always scan the document at good quality.

If the scanned image is of good quality it is easy to extract text from images. To extract text from images we need OCR (Optical Character Recognition) technique which is use to extract text from images it will helps us to extract texts which are present in images.

If the required keyword is present in those images, then it will check how many times the same keyword is occur in that image and as per the number of time the keyword is found in that image it will provide the priority to those images, and it will arrange all these images which contain the same keyword maximum times.

II. Problem Definition

. The most prominent challenge faced by search engine is to search an image which has the content keyword which user wants. The search engine can't give the images which have that given keyword.

The result of the image is processed by their name, if the given keyword is in the name of the image. To extract the text from image is not an easy job to do. There is image processing technique is use to extract text from image.

From huge database selecting such images which contain only the given keyword present on those images.

III. Objectives of the study

.

Use OCR technique which focused on extract text from images.

OCR is used when recreating a similar document in paper as a document in electronic form takes more time.

The converted text files take less space than the original image file and can be indexed. Hence the use of OCR adds an advantage to the user who had to deal with conversion of great amount of paper works in to electronic form.

At first, text regions are extracted and skew corrected. Then, these regions are binarized and segmented into lines and characters. Characters are passed into the recognition module.

Using extracted text, searching of keyword is become an easy to perform. Where extracted text is convert in to the text file.

Search method is search keyword in the text file. Then it is become an easy to find the string keyword.

After searching method result is found in the sorted format, which is on the basis of the image how many times keyword is found.

IV. Research Design

Focused searching and mine image from files and folder algorithm

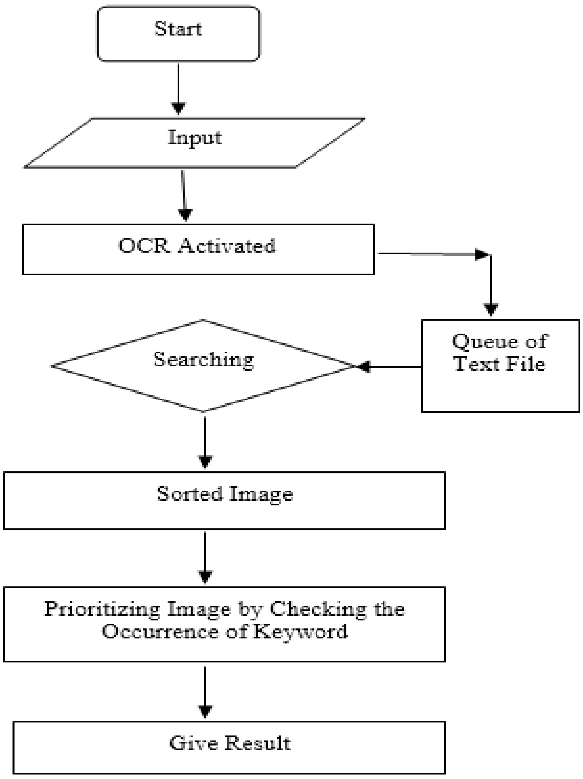


Fig 1. Project Implementation Flow

V. Methodology

The first step is to search the image in database (or file storage). This section involves to find the require keyword in that particular image. Next extract each image to find the keyword is present in the image or not. For that An Optical character recognition system used, which consists of Text Extraction, Skew Correction, Binarization, segmentation, recognition, and post processing stages.

A. Text Region Extraction-

Image are given as input from that image text is extracted by selecting area over text and that present text on image is extracted The input of this module is an image and output is a set of rectangles surrounding each text region of the input image. The text detection module applies some pre-processing algorithms on the input image, such as greyscale transformation, filtering, binarization, decomposition, scaling etc. After the pre-processing steps, a classification algorithm separates the texts form non-texts.

B. Skew Correction-

An image which is captured from Camera is very often suffered from skew and perspective distortion. Skew and perspective distortion occurs due to unparallel axes or planes at the time of capturing the image. There are two types of pixels in every text region – dark and gray. The dark pixels constitute the texts and the gray pixels are background around the texts. The text which is not in proper from or inclined to the horizontal line is corrected. By calculating its height and distance according to the inclined image it is set to the proper on horizontal line.

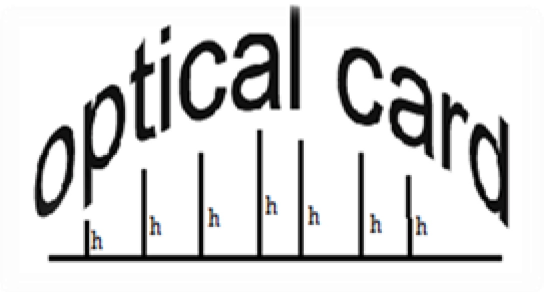


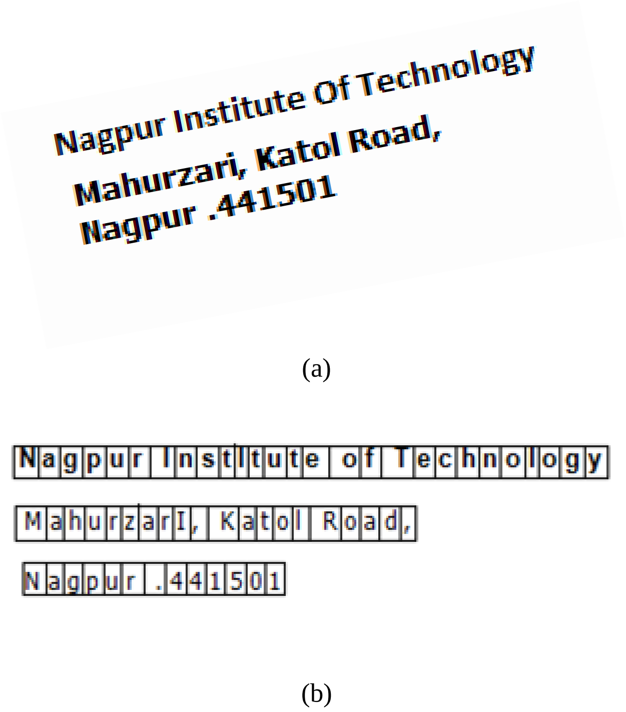
Fig 2: Calculation of skew angle from bottom profile of a text region

C. Binarization-

The binarization is a process of separating the background and foreground text. A skew corrected text region is binarized using a simple yet efficient binarization technique developed by us before segmenting it. After document binarization a top-down segmentation approach is applied. First lines of the documents are detected, then words are extracted and finally words are segmented in characters.

D. Text Region Segmentation-

After Binarization process, the selected portion from image is extracted and that extracted portion is grouped in the blocks. At first, all possible line segments are notified by commencement the own values. The commencement is selected so as to grant over-segmentation. Text line boundaries are referred by the position of the character which is referred. These rectangular boundaries are set on the all characters. Because of this segmentation reorganization is become easy. And then there every character is looks individual to the reorganization method. The process of reorganization method is become more efficient.



(a)

(b)

Fig. 3 . Skew correction and segmentation of text regions,
(a) An extracted text region,
(b) Characters segmented from de-skewed text region

E. Character Recognition-

The character recognition technique help to identify the text, which available in the image. Because of the segmentation every character is set as individual. The segmentation is gives the boundary rectangle on every character then every character is recognise easily.

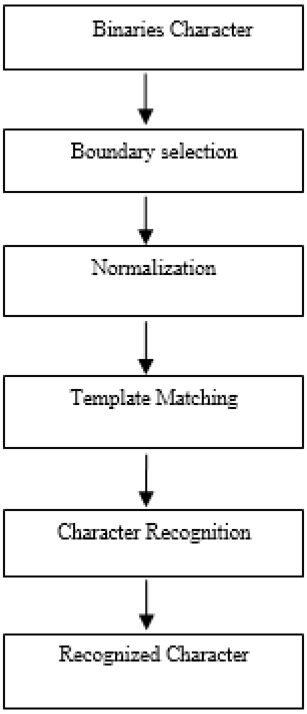


Fig. 4. Block diagram of the character recognition module

The serial steps required to classify an individual binarized character. After resizing the paradigm by its segmentation box, it is normalized to a model dimension, 48x48. Among the 256 ASCII characters, only 94 are use in image, but only 73 characters are frequently use in the documentation. There are 26 capital letters, 26 small letters, 10 numeric digits and 11 special characters. This all 73 classes ares shows in below table.

A	B	C	D	E	F	G	H	I	J
K	L	M	N	O	P	Q	R	S	T
U	V	W	X	Y	Z	a	b	c	D
e	f	G	h	i	j	k	l	m	N
o	p	Q	r	s	t	u	v	w	X
y	z	1	2	3	4	5	6	7	8
9	0	#	@	&	*	(	)	.	,
+	-	:							

Table : 73 Classes Recognition Problem.

VI. Conclusion

In this system we are developing a search engine to find the image by providing keyword the text is mined from image. The user enter the keyword for search, if that keyword is present in image which is available in database then the text is mined from image with the help of extraction technique and the searching and sorting algorithm is used to gives the result.The basic procedure is to identify “text” versus “background” pixels then extract individual symbol.An Optical character recognition system used, which consists of Text Extraction, Skew Correction, Binarization, segmentation, recognition, and post processing stages

REFERENCES

[1]G.Vamvakas, B.Gatos, N. Stamatopoulos, and S.J.Perantonis "A Complete Optical Character Recognition Methodology for Historical Documents" Computational Intelligence Laboratory, Institute of Informatics and Telecommunications, National Center for Scientific Research.

[2]Ayatullah Faruk Mollah, Nabamita Majumder, Subhadip Basuand Mita Nasipuri "Design of an Optical Character Recognition System for Camera-based Handheld Devices (July 2011)"Department of Computer Science and Engineering.

[3] A. F. Mollah, S. Basu, N. Das, R. Sarkar, M. Nasipuri, M. Kundu, "A Fast Skew Correction Technique for Camera Captured Business Card Images", Proc. of IEEE INDICON-2009

[4]http://code.google.com/p/tesseract-ocr

[5] A. F. Mollah, S. Basu, M. Nasipuri, "Segmentation of Camera Captured Business Card Images for Mobile Devices", Int'l J. of Computer Science and Applications, 1(1), pp. 33-37, June 2010.

[6] MdZahidulIslam and Amit Kumar Mondaly "Towards a Standard Bangla PhotoOCR: Text Detection and Localization" Computer Science and Engineering Discipline Khulna University, IEEE December-2014.