

RESEARCH ARTICLE



CASE STUDY: A FRESH APPROACH FOR OCR (OPTICAL CHARACTER RECOGNITION)

S.ARIF ABDUL RAHUMAN<sup>1</sup> , DR.J.VEEAPPAN<sup>2</sup>

<sup>1</sup> Prof/CSE , M.E.T Engineering College , Tamilnadu , India

<sup>2</sup> Prof/ECE , Sethu Institute of Technology , Tamilnadu , India

Article Received: 23/12/2013

Article Revised on: 09/01/2014

Article Accepted on:12/01/2014



Scan QR Code for  
More details

**ABSTRACT**

In this modern era, everybody wants the information to be stored, retrieved and processed as fast as possible. Paper based Documentation has become obsolete these days, all the information are captured in digital (electronic data) format and stored in secure place. The methodology discussed in the present paper reduces the storage cost , ensures the security and increases the data accessibility. Most of the organizations have already stepped into the digital world. Here, the concern is, their past information is yet to be migrated from image or paper based document into digital text based documents. So, every organization expects some progress in such kind of conversion or migration or some modern methodology which serves their purpose.

Such a concept became an emerging technology in image processing as an Optical Character Recognition methodology.

Here, we present a brief analysis to identify, understand new solution, evaluate and execute this methodology. Also we propose to overcome the issues which have been identified in the existing Optical Character Recognition Solutions.

[The present work proposes solution only for typed text characters, not for hand – written text ]

**INTRODUCTION**

Optical Character Recognition henceforth will be termed as OCR, is a process of extracting the relevant information from raw data. This has been defined especially in text domain, as “Sentence or Word or Character extraction from an image source. In order to preserve the content from high volume of Image data into light weight text data in an electronic format ”. The flow diagram ( Image 1.1 ) depicts the process. There are six steps involved here, two among them need user intervention ( to feed in the input image and to store the result in preferred document format). Other four steps are explained in detail below,

**a. Data Feed-In:**

Scanned document(s), as an image, are provided to the system via scanners or by other means. Here, the input data can be single file / multiple files.

**b. Pre – Processing:**

This is not maintained as mandatory process in some existing solutions, but will improve the output quality.

Different pre-processing techniques have been used like Noise removal ( Salt-pepper noise removal, ink plot removal ,etc.) Zero degree positioning (Scanned document might be tilted ; increasing the ambiguity, with poor results, because text direction always matters in OCR processing. This pre – processing is achieved through the basic image processing techniques.

**c. Character Recognition:**

After the pre-processing, grouped text / words are segregated as a single character and processed. Each character is recognized based on the technology used. There are different approaches to recognize the characters; most commonly, Pattern Based approach is used, which provides more accuracy than other approaches.

**d. Text – Optimization:**

Using the pre-defined patterns the extracted character images are compared. The matched image will be converted as text corresponding to the matched pattern structure. Some tools preserve the text’s position and structure (size, thickness, fonts, etc.) as well.

Recognized text positioned as text and non-recognized images will be ignored or displayed as it is. Almost all tools consider these non-recognized images as noises / unnecessary data and ignore while preparing the final result. So, the non-recognized character images will be replaced by blank space. [Here also position and structure are preserved in some tools like recognized text].

Successful completion of these steps will yield the final result, which can be stored for future purpose. Here, the fed data and final result will be having some content but the resultant document can be edited and made easily manageable which cannot be done in fed data.

**2. OCR TOOL AND USAGE OVERVIEW**

Many domains make use of OCR. Earlier, it was used only in government organizations and also only for specific characters, like numeric character identification, to identify the numerals in the number plates. Later its scope extended to alpha numeric then it further extended for local languages too.

Some of the domains and its usages are provided below,

a. Transportation

Road and Traffic control, Rash drive or Traffic violation identification.

b. Information Technology (BPOs)

Image document to voiced / text based document conversion.

c. Banking

Account and Chelan number identification from cheques / DDs.

d. Scientific / Robotics

Character identification and voice conversion from Text ( like from Newspapers / books ) for physically challenged people ( This is still under processing ).

Also, based on the usage / scope of conversion, plenty of OCR tools available in market, each focuses on particular area and provides the required result. Here, it’s being more or less domain specific.

For e.g. If the tool has been developed for the media domain, it cannot be used for scientific domain. Because, in media it may concentrate more on text, but in scientific domain focus should be on scientific fields ( like Formula, tables and matrix etc. ) too.

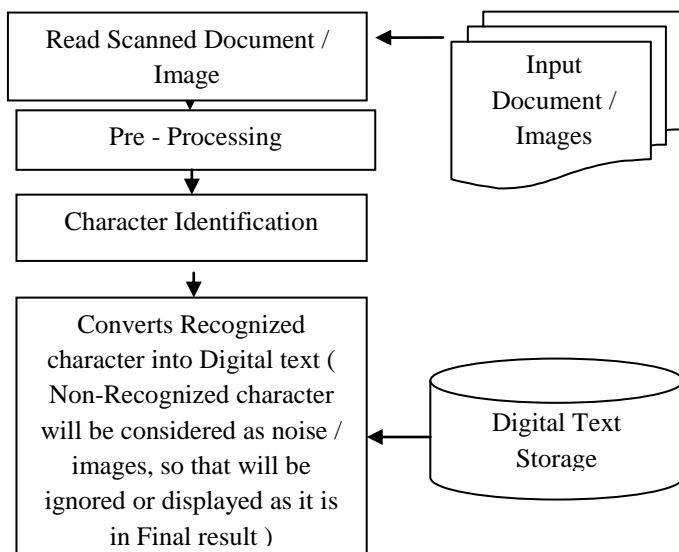


Image1.1 ( OCR Process Flow)

3. ANALYSIS ON OCR

Comparative study has been conducted on different OCR tools which are available in market. In order to conduct this test two major areas have been focused. One is based on its name / id ( how it is pronounced ) and another one is based on its configuration( how it is written )

By name / Id: which is further divided and grouped as,

a . Textual: Alpha, Numeric, Punctuation, Special Symbol – one( from key board ), Special Symbols – two( from scientific formulas )

b. Non – Textual: Images, tables, etc.

By Configuration: which is sub divided into,

a. Font: Written technique

b. Face: Bold, Italic

c. Size: Size of the text

Using the combination of Name / Id and configuration, different set of samples have been prepared. All the samples are fed into different tools and the results are captured and tabulated here,

Combos	Textual					Non - Textual		Font		Face		Size	
	Alpha	Num	Punc	Spl 1	Spl 2	Image	Table	1	2	Bold	Italic	1	2
Com-1	X		X	X				X				X	X
Com-2		X	X	X			X	X	X		X		X
Com-3	X	X	X	X		X		X	X	X		X	
Com-4	X		X		X	X	X			X	X	X	X
Com-5		X	X		X			X	X				X
Com-6	X	X	X		X		X	X	X		X	X	
Com-7	X	X	X	X	X	X			X	X		X	X
Com-8	X			X		X	X		X	X	X		X
Com-9		X		X				X				X	
Com-A	X	X		X			X	X			X	X	X
Com-B	X				X	X			X	X			X
Com-C		X			X	X				X	X	X	
Com-D	X	X			X			X	X			X	X
Com-E	X	X			X		X	X			X		X
Com-F	X	X				X	X		X	X		X	
Com-0	X	X	X	X	X	X	X	X	X	X	X	X	X

“X” / “” | each column / row represents the presence / absence of that particular text / character.

E.g. “X” under column “Alpha” represents the alphabets that are present in the Scanned input image, “” under column “Image” indicates that there will be no image( drawing / photograph ) present in the scanned input image.

Based on the above mentioned combinations there were 16 images prepared. All the 16 images were fed into different tools and results displayed in the table shown below as percentage of accuracy.

Combination(s)	Accuracy Across Different Tools				
	Tool - 1	Tool – 2	Tool – 3	Tool – 4	Tool - 5
Combination-1	0.5	0.5	0.5	0.5	0.5
Combination-2	0.3	0.4	0.6	0.3	0.2
Combination-3	0.4	0.5	0.4	0.3	0.4
Combination-4	0.3	0.3	0.3	0.3	0.3
Combination-5	0.5	0.5	0.6	0.2	0.3
Combination-6	0.3	0.4	0.5	0.3	0.2
Combination-7	0.5	0.5	0.5	0.5	0.5
Combination-8	0.2	0.3	0.4	0.1	0.2
Combination-9	0.5	0.5	0.6	0.5	0.5
Combination-A	0.2	0.3	0.3	0.3	0.3
Combination-B	0.4	0.4	0.2	0.4	0.4
Combination-C	0.3	0.3	0.3	0.3	0.3

Combination-D	0.4	0.4	0.5	0.2	0.4
Combination-E	0.4	0.4	0.5	0.2	0.4
Combination-F	0.3	0.3	0.3	0.3	0.3
Combination-0	0.3	0.3	0.5	0.3	0.3
Output Image	IMTO - 1	IMTO - 2	IMTO - 3	IMTO - 4	IMTO - 5

In the following cases the sample input images contain all the possible combinations like Alphanumeric, Alphabet, Numerals, Punctuation, Special characters and images with different Font Sizes.

There are 52(fifty two) characters in Alphabet (a-z, A-Z); there are 10( ten) characters in Numerals(0-9).

Following table depicts the same,

Image Type	Image	Remarks
Alphabet	H <sup>i</sup> D <sup>G</sup> P <sup>Q</sup> g <sup>L</sup> e <sup>O</sup> X <sup>h</sup> Y <sup>i</sup> T <sup>n</sup> j <sup>W</sup> q <sup>r</sup> y <sup>C</sup> r <sup>w</sup> s <sup>f</sup> b <sup>z</sup> F <sup>N</sup> c <sup>V</sup> K <sup>d</sup> o <sup>o</sup> k <sup>a</sup> I <sup>AS</sup> x <sup>m</sup> t <sup>Z</sup> R <sup>v</sup> E M u U	a-z & A-Z characters are available in this image
Numeric	9 5 <sup>8</sup> <sub>1</sub> 6 0 4 3 7 <sup>2</sup>	0-9 characters are available in this image

Output Images:

IMTO - 1

Ima-e TV11c 11/ lgg Remarks Alphabet H 1 O G Q O Oh P gp r i T ,... , q y r C w , f b <sup>o</sup> ' - 2c> K d <sup>o</sup> aN, AS n., x t u j ' 1.1 V a-z & A-Z characters are available in this image Numeric 901 8 3 00 97107 7 0- 9 characters are available in this image
---

IMTO - 2

This is a sample input image. This contains all the possible Combinations like Alphabets, Numerals, Punctuation, Special Characters and Images with Font phases. There are 52 characters in Alphabet( a-z & A-Z); there are 10 characters in Numerals(0-9). Following Image depicts the same, .....l¥,,.....e..... H D 6 Pg Q 1 O X v E W H P I T * I A x&AeZ dnnzaensaneavanlalzk Mm a v r gwsf ' mmmmug Be _ zc> dO Ka ,A 5 mx Np, V 9 : 15 0-9clul : aeuale: vallal : kmLIns Nmnenc 3 m Hz e
---

**IMTO – 3**

Im"g T		
Alpb	<p>H<sup>I</sup> D<sup>G</sup>                  P<sup>g</sup> Q<sup>q</sup> L<sup>e</sup> O<sup>o</sup>                  X<sup>h</sup> Y<sup>i</sup> T<sup>n</sup> j<sup>j</sup> W<sup>w</sup>                  q<sup>q</sup> y<sup>y</sup> r<sup>r</sup> C<sup>c</sup> K<sup>k</sup> d<sup>d</sup> w<sup>w</sup> o<sup>o</sup> s<sup>s</sup> f<sup>f</sup>                  b<sup>z</sup> zc&gt;                  k<sup>a</sup> l<sup>AS</sup> m<sup>x</sup>                  t, . su v</p>	
	<p>39                  Z                  7</p>	<p>9 n<sub>1</sub><sup>8</sup>                  o</p>

**IMTO – 4**

hi  
 l  
 ll D G  
 Pg Q L O  
 X h Y . 9 .W  
 p l T " l &AZ characters are available  
 Hab t r C sf a'Zi. . .  
 ALp e q y K W mtruslimage  
 Bl u. Zc> d O  
 T N an  
 l l l 5": V  
 9" 'l 8  
  
 . 1 ° 0-9 characters are available mthrs  
 Numenc 3 9 .  
 v 7 Z mge

**IMTO – 5**

.....1¥,.....e.....  
 H D 6  
 Pg Q l O  
 X v 9 W  
 h p l T \* l A x&AeZ dnnzaensaneavanlalzk  
 mm a v r gwsf ' mdmmug  
 be \_ zc> dO  
 Ka ,A 5 mx  
 . Np, V  
 9 ::{e9dn | : aa : aleavall: l : kmLlm  
 em 3 M  
 Hz e

## 5. CONCLUSION

The present review leads to the conclusion that there are plenty of enhancement points left open. Current Optical Recognition process is yet to be completed with modern methodology to ensure the desired result, as the result achieved over this thesis does not seem to meet the desired level. Results from different applications are not so consistent with original input. Hence, we are proposing a new solution with new methodology, which will address the issues discussed here. This solution will be discussed in an another document.

## 6. REFERENCE

Following are the references from web resources which have been used to pursue the present analysis,

1. [http://en.wikipedia.org/wiki/Optical\\_character\\_recognition](http://en.wikipedia.org/wiki/Optical_character_recognition)
2. <http://www.onlineocr.net/>
3. <http://www.onlineocr.net/default.aspx>
4. <http://www.ocr.org.uk/>
5. <http://www.free-ocr.com/>
6. <http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=240312>
7. <http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=156468&userType=inst>
8. [http://ieeexplore.ieee.org/xpls/abs\\_all.jsp?arnumber=565084&tag=1](http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=565084&tag=1)
9. <http://arxiv.org/ftp/arxiv/papers/1109/1109.3317.pdf>
10. <http://robotics.usc.edu/publications/media/uploads/pubs/635.pdf>
11. <http://free-online-ocr.com/>
12. <http://www.engpaper.com/free-research-paper-optical-character-recognition-ocr.htm>
13. <http://www.gobokee.net/ieee-papers-ocr/>
14. <http://technav.ieee.org/tag/7695/optical-character-recognition-software>
15. [http://en.wikipedia.org/wiki/Optical\\_character\\_recognition](http://en.wikipedia.org/wiki/Optical_character_recognition)
16. <http://www.i2ocr.com/>
17. <http://www.ocr-extract.com/result.html?extractUUID=3b42daba-bfbf-40ca-b743-f440945f8838>
18. <http://www.verypdf.com/online/ocr-converter.php>
19. <http://www.freewaregenius.com/how-to-extract-text-from-images-a-comparison-of-free-ocr-tools/>
20. <http://gigaom.com/2011/04/11/3-free-online-ocr-tools-for-your-paperless-office/>
21. <http://www.to-text.net/>
22. <http://www.nemopdf.com/knowledge/top-4-free-online-ocr-tools.html>
23. <http://www.labnol.org/software/convert-images-to-text-with-ocr/17418/>