International Journal of Multidisciplinary Approach and Studies

ISSN NO:: 2348 – 537X

A REVIEW ON CHARACTER RETRIEVAL TO SUPPORT READING DOCUMENTS

Aman Parkash Singh * & Mr. Sandeep Kaushal **

[*] M.Tech Scholar, Department of Electronics and Communication Engineering, ACET, Amritsar, Punjab, India,

[**]Associate Professor, Department of Electronics and Communication Engineering, ACET, Amritsar, Punjab,,India

www.ijmas.com

ABSTRACT:

Information spotting in scanned document images is a very challenging task. Character recognition has been studied from the past several decades and is still a demanding research topic in the field of pattern recognition and image processing, however their performance will be significantly impaired when the image of the character is partially blocked or smudged. Such missing information does not hinder the human perception because we predict the missing part based on the word level and sentence level context of the character. This paper provides a comprehensive review of existing works in character retrieval based on different methodology.

Keywords:-*Fuzzy System, HMM, Image Processing, , Non-linear normalization.*

I. INTRODUCTION

Text in an image contains meaningful and useful information which can be used to fully understand the contents of the images. Text extraction from images play an important role in document analysis, document retrieval, blind and visually impaired users etc. A document image contains various information such as line drawings and sketches. They are developed by scanning journals historical document images, degraded document images. In real life text exist in many forms other than its ASCII representation. There include printed written texts. There are many occasions when only the scanned or photographed image of text is available for computer processing. While the machine reading system bridges the gap between natural language and artificial intelligence, another bridge has to be constructed to link the natural state of texts to its unique encoding that can be understood by computers. Review of text extraction and character recognition. An image is an array, or a matrix of square pixels arranged in column and rows. In an (8-bit) greyscale image is what people normally called a black and white image but the name emphasizes that such an image will also include many shades of gray. A normal greyscale image has 8-bit colour depth = 256 gray scale. A true colour image has 24-bit colour depth= 8x8x8 bits=256x256x256 colours.

II. FUZZY IMAGE PROCESSING

Mohanad Alata [1] suggested fuzzy image processing for text detection and character recognition. In his paper they deal with image which means that 256 different colours. This makes dealing with each single colour harder and they are also deal with colour because if the colour of the text is almost as that of the background it will be last. In his work they first tried to reduce the colour that is less than 60 pixels. The presented approach will take each colour alone and treat it as ON and the other as OFF. This makes the merge between the text and the background unlikely to happen. Also there is no negative text in the whole image because the negative text to a colour will be normal text for another. They used font size between 7-29 point for the three type of font these are Verdana, Arial and Lucida console. These types were chosen because the characters have low variance and there is less redundancy in the single characters.

The first step in his approach is to decreasing the number of colors in the image by ignoring colors which have number of pixel below 60 pixels, and by decreasing the level of color will be treated as a binary image by considering the color ON and the others to be OFF.In each binary image the connected components will be detected and segmented to classify it as a character or noise. The restrictions are:

There are at least three objects in a single line having the same color.

The object properties match the threshold points.

The object area is not below 20 pixels.

The height and width are not above 45, 40 pixel, respectively.

In this work, a connected component method was used to establish the detection algorithm and the recognition algorithm. A computer program was developed to implement the algorithm for three types of fonts and sizes within 17-29 points.

www.ijmas.com

Fuzzy System

Both fuzzy systems have been developed using Matlab Fuzzy toolbox. The systems are based on Mamdani fuzzy approach. The first task is to define the inputs and the outputs of the fuzzy system. This stage depends on expert decision. For the first fuzzy system the inputs are in Figure1.

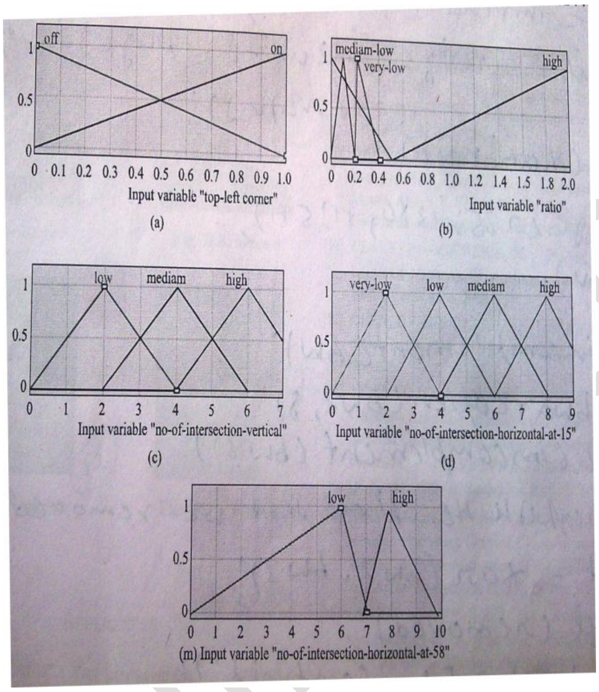

Figure.1. (a) input 1(upper-left corner), (b) input 2 (upper-middle corner), (c) input 3(upper-right corner), (d) input 4(middle-right corner), (e) input 5(lower-right corner), (f) input 6(lower-middle corner). [1]

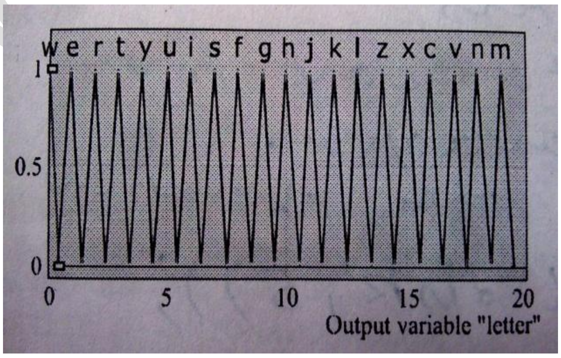

Figure.2. The output of the first system.[1]

www.ijmas.com

The output membership of the first system is in Figure2.

In Fuzzy system design the expert will identify the commands (rules) that will model the relationship between the inputs and the output. Some rules listed here for the fist system.

Upper left → UL Lower left→LL

Lower middle→ LM Lower right→LR

Upper middle→UM Upper right→UR

Middle left→ML Middle middle→MM

Middle right→MR Ratio length→R

Position hole x→X Position hole y→Y

No. of line intersection vertically→V

No. of line intersection horizontally 25%→ H25

No. of line intersection horizontally 75%→H75

III. PERCEPTION-PREDICTION MODEL FOR CONTEXT AWARE TEXT RECOGNITION

Qinru Qiu, Qing Wu, and Richard Linderman[2] suggested perception-prediction model for context aware text Recognition on a heterogeneous many core Platform. In his paper they present a unified perception-prediction framework that combines the algorithms of neural networks and confabulation. The framework uses neural network, models for pattern recognition from raw input signal, and confabulation models for abstract-level recognition and perdition functionalities. The prototype of a context aware Intelligence Text Recognition System(ITRS) that mimics the human information processing procedure. The ITRS system learns from what has been read and, based on the obtained knowledge; it forms anticipations and predicts the next input image or the missing part of the current image. Such anticipation helps the system to deal with all kinds of noise that may occur during recognition.

www.ijmas.com

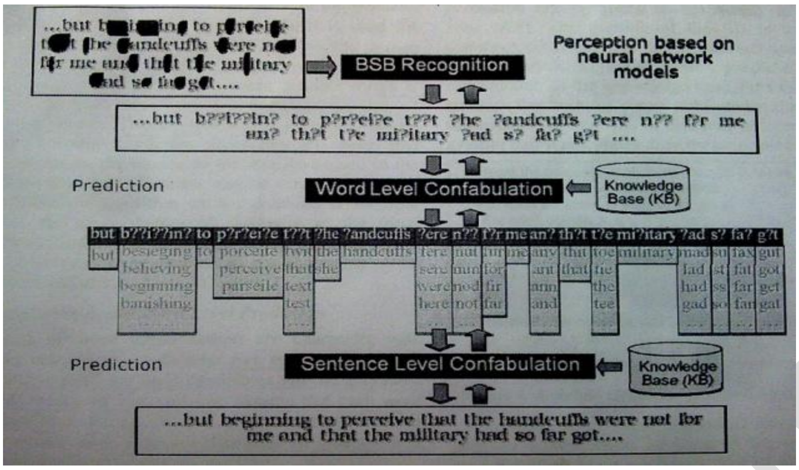
...but b??i??in? to p?r?ei?e t??t ?he ?andcuffs ?ere n?? f?r me an? th?t t?e mi?itary ?ad s? fa? g?t ....
BSB Recognition
Perception based on neural network models
Prediction
Word Level Confabulation
Knowledge Base (KB)
but b??i??in? to p?r?ei?e t??t ?he ?andcuffs ?ere n?? f?r me an? th?t t?e mi?itary ?ad s? fa? g?t
Prediction
Sentence Level Confabulation
Knowledge Base (KB)
...but beginning to perceive that the handcuffs were not for me and that the military had so far got....

Figure.3. The overall architecture of the models and algorithmic flow.[2]

The ITRS is divided into 3 layers as shown in Figure 3. The input of the system is the text image. The first layer is character recognition software based on BSB models. All potential candidates will be reported as the BSB results. Using the racing model, if there is noise in the image or the image is partially damaged; multiple matching patterns will be found. For example a horizontal scratch will make the letter 'T' look like letter 'F'. In this case we have ambiguous information. The ambiguity can be removed by considering the word level and sentence level context, which is achieved in the second and third layer where word and sentence recognitions are performed using cogent confabulation models. The models fill in the missing character in a word and missing words in a sentence. The three layers works cooperatively. The BSB layer performs the word recognition and it sends the potential letter candidates to the word level confabulation based on those letter candidates and sends this information to the sentence recognition layer. There could be feedback paths to word level or send word confabulation result back to character level. In the Figure 3 the BSB algorithm recognizes text images with its best effort. The word level confabulation provides all possible words that associate with the recognized characters while the sentence level confabulation finds the combination among those words that gives the most meaningful sentence.

IV. SIMILARITY EVALUATION AND SHAPE FEATURE EXTRAACTION

Akihito Kitadai and Masaki Nakagawa [3] they present similarity evaluation method for character patterns with missing shape parts they worked with non-linear normalization for

www.ijmas.com

such patterns, and modifies the templates for each trial of the retrieval efficiently. They used kanji character patterns from the Japanese historical documents called mokkans, they also present a simple implementation of gradient feature extraction to compare the chain code feature with the gradient feature in the retrieval. As the result, the gradient feature works better than the chain code feature. One of the serious problems for the CPR (Character Pattern retrieval) is missing shape parts of character patterns. On the historical mokkans, decayed and discoloured surfaces of the wooden tablets and tarnished ink generate lots of missing shape parts in the character patterns. Therefore they used robust techniques of similarity evaluation for the CPR. Another problem is unstable shape features of character patterns. Even if we employ accurate image processing, the patterns contain noise coming from degraded documents. They used Non-linear normalization using histograms of shape features in X and Y coordinates performs well in character patterns recognition.

www.ijmas.com

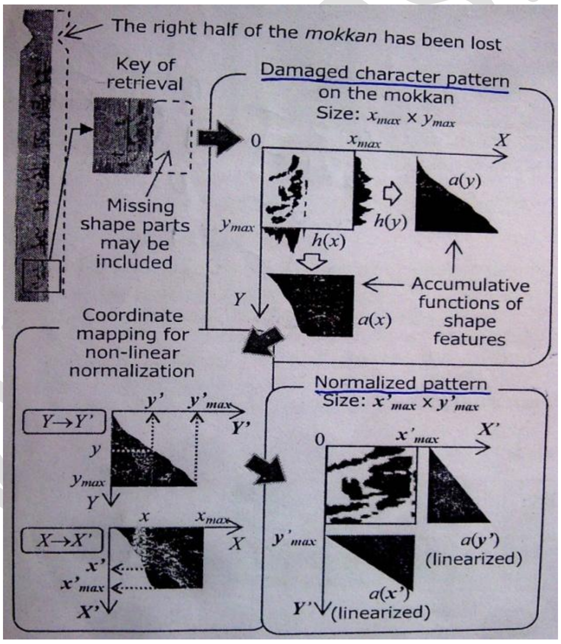
The right half of the mokkan has been lost
Key of retrieval
Damaged character pattern on the mokkan
Size: $x_{max} \times y_{max}$
0
x_{max}
X
$a(y)$
$h(y)$
y_{max}
$h(x)$
Missing shape parts may be included
Y
$a(x)$
Accumulative functions of shape features
Coordinate mapping for non-linear normalization
Normalized pattern
Size: $x'_{max} \times y'_{max}$
Y→Y'
y'
y'_{max}
Y'
y
y_{max}
Y
X→X'
x
x_{max}
X
x'
x'_{max}
X'
0
x'_{max}
X'
y'_{max}
$a(y')$ (linearized)
$a(x')$
Y' (linearized)

Figure.4. Problem of non-linear normalization.[3]

However deformation by the normalization is too much for the keys of character pattern retrieval with missing shape parts. Figure4 shows the problem h(x) and h(y) are the histograms, a(x) and a(y) are the accumulative functions of h(x) and h(y), a(x’) and a(y’) are the linearized a(x) and a(y). To manage the deformation, his normalization method employs the gray-zones painted by the archaeologist and historians with digital pointing devices.

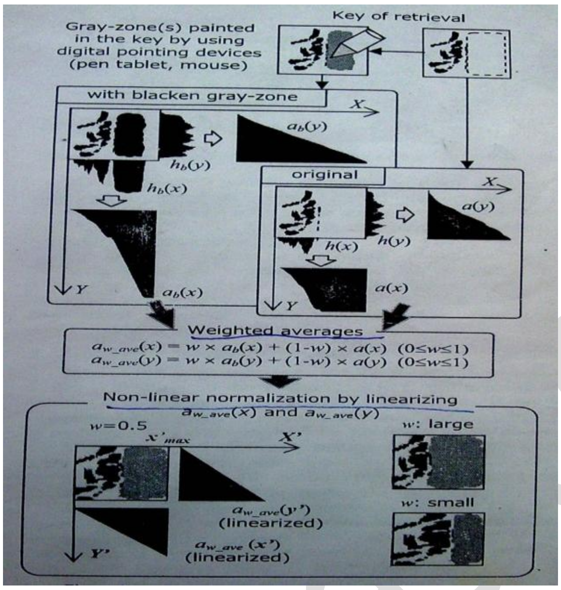
Gray-zone(s) painted in the key by using digital pointing devices (pen tablet, mouse)
Key of retrieval
with blacken gray-zone
X
$a_b(y)$
$h_b(y)$
$h_b(x)$
Y
$a_b(x)$
original
X
$a(y)$
$h(x)$
$h(y)$
$a(x)$
Y
Weighted averages
$a_{w\_ave}(x) = w \times a_b(x) + (1-w) \times a(x)$ $(0 \leq w \leq 1)$
$a_{w\_ave}(y) = w \times a_b(y) + (1-w) \times a(y)$ $(0 \leq w \leq 1)$
Non-linear normalization by linearizing $a_{w\_ave}(x)$ and $a_{w\_ave}(y)$
w=0.5
x'max
X'
$a_{w\_ave}(y')$ (linearized)
$a_{w\_ave}(x')$ (linearized)
Y'
w: large
w: small

Figure.5. Interactive non-linear normalization.[3]

They show the flow of the method in Figure 5 $h_b(x)$ and $h_b(y)$. In this method, they linearize the weighted averages of the accumulative function in each direction: $a_{w\_ave}(x)$ and $a_{w\_ave}(y)$.

Creating feature matrix

They employ chain code feature as the shape feature of character patterns. By scanning the pixels in the non-linear normalization character patterns with the following 3x3 pixel-patterns (Figure 6).

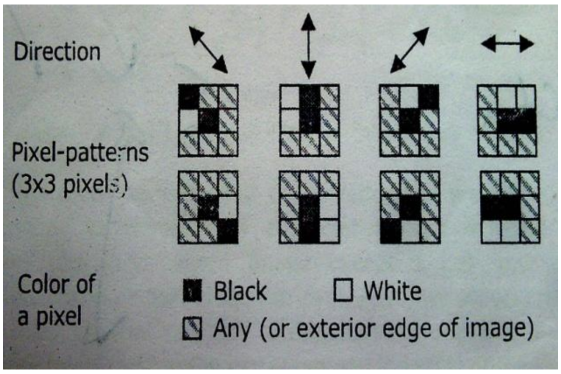

Figure.6. Pixel-patterns for 4-directional chaincode feature.[3]

www.ijmas.com

They obtained 4-directional chain code features. In the sampling process of CPR method, they accumulate the shape features with multiple Gaussian functions (Figure. 7). Each of the Gaussian function has a unique centre point that is an intersection point of the grid, and amplifies the features close to the point. Gaussian functions as G (h,v). The (h,v) is the center point in which h, v x shows the column number and y shows the row number of the grid. Using 8x8 equally-separated grid for the Gaussian functions to obtain 64 accumulated shape features for each direction. We present each of the accumulated shape features F(d, h, v). Finally the dimension of feature matrix becomes 4x64 for each character pattern. The feature matrices of the character patterns images in the archives of the mokkans become the templates of CPR . For the similarity evaluation between each of the templates and a key, they employ the negative value of the city block distance.

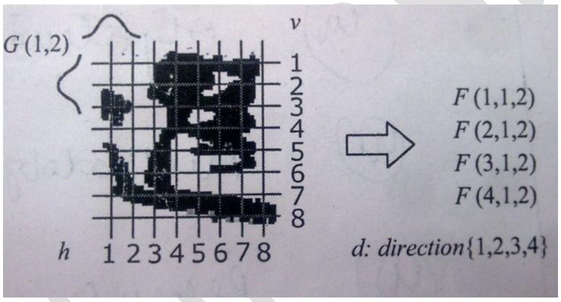

Figure.7. Sampled features: F(d,h,w) with a Gaussian function G(1,2). [3]

Problem in similarity evaluation

Even if the gray-zone correct the X and Y coordinates of shape features in the normalized key, the missing shape features are not sampled from the gray zone. It produces unjust distance between the key and the genuine template(s) of retrieval. They proposed average feature(AF) method. The AF method does not reconstruct the templates. They propose another method to absorb the unjust distance efficiently. In his paper they call it template modification(TM) method. The TM method does not reconstruct the feature matrices of the templates or inject any shape feature into the gray-zone. In each calculation for similarity efficiently, then TM method just reduces each F (d,x,y) in the feature matrices of the templates by considering the gray-zone in the normalized key.

www.ijmas.com

For each pixel in the normalized key with gray-zone, the TM method gives a score $S_{gray.}$

$S_{gray=}$ a, for pixels in gray- zone.

0, for pixels not in gray-zone.

The value ‘a’ is constant and not equal to 0. Also , the TM method gives the other score s=a to every pixel in the normalized key. The TM method accumulates the S_{gray} and S with the same Gaussian functions G(h,v). We present each accumulated result as S_{gray}(h,v) and S(h,v). Finally the TM method replaces the F(d,h,v) in the feature matrix of the templates by the F’(d,h,v) as presented in the equation below.

$F'(d,h,v)= F(d,h,v)x\{1-S_{gray}(h,v)/ S(h,v)\}$

They made 10 mask images to generate missing shapes parts or gray-zones on the character patteren images artificially. Such managed missing shape parts and gray-zones are suitable for quantitative evaluations of CPR. The hit rate without the mask image was 81.7%. Also, the hit-rate was 44.2% when they used when they used masked images with the w=0(missing shape parts).

V. SHAPE AND SEQUENCE COMPARISON

Khurram Khurshid and Claudie Faure [4] proposed a word spotting method for scanned documents in order to find the word imaged that are similar to a query word, without assuming a correct segmentation of the words in such documents. Aim is to facilitate the information search by spotting the different instances of a given query word in documents. Word matching has to handle local shape distortions as well as inexact segmentation of the words to compare. This is achieved by coupling local shape comparison at sub-pattern level and string comparison at word level.

In his proposed method, language independent features are preferred to captured the information of shape without any prior assumption on the writing system used in the document. Two levels of representation are defined to evaluate the matching of word patterns as a combination of local similarities between sub-patterns. As only Roman writing system are encountered in the processed data, the sub-patterns are defined at character level and their spatial organization in sequences is in agreement with the way printed words are composed with contiguous character type sorts.

www.ijmas.com

Shape matching is performed at local level with a DWT comparison of S-characters. At word level, the comparison of the S-character sequences is performed with a modified Edit distance or a linear string matching process.

The connected components of the binarized image are processed to obtain a sequence of s-characters for each word. Sequence of feature vectors are computed for the S-characters to achieve word indexing.

The set of feature adopted for his study is

F1: the vertical projection calculate on the gray level image.

F2: the upper profile.

F3: the lower profiles.

F4: ink/non-ink transition in an image column.

F5: the vertical histogram i.e. no of black pixels in each column.

F6: the transition status of the mid row pixel along a horizontal line

They implement multi-step comparison process to retrieve the word similar to the query. The aim of the first step is to filter the number of words to be compared with the query. A coarse criterion rapidly eliminates a large amount of words from the candidates to be compared without eliminating the relevant words.

Indexing process is the computation of the word as it is time consuming process, document image indexing is done beforehand to allow a rapid information search. A file is associated with each document image and contains the coordinates of each word in the page, the number and order list of the S-characters in the word, their positions and their computed feature vectors. This work provides a thorough examination of segmentation based retrieval techniques for historical document images. In his system, it allows queries either in the form of a word image or as an ASCII text.

www.ijmas.com

VI. HIDDEN-MARKOV-MODEL BASED SEGMENTATION

Aman Parkash Singh and Sandeep Kaushal [5] proposed Hidden Markove Model(HMM) for retrieval of degraded character in the documents. Image Processing on character are very intrusting and challenging task we are studied various method which are used for retrieval of degraded or missing character. They used some documents which are degraded and perform various operations to filter the image and used for retrieval using HMM. They convert the RGB image into gray and binaries the image and also apply median filter to reduce the noise. After this binarization process is applied to obtain character in each candidate region. After the candidate region have been obtained the adjacent region are first grouped. When the objet in each region are extracted their position in the original image can also be obtained. After this they find number of character in a word. In his work this is the main challenging task because if last character of a word and the first character of the second word in the same line is missing then it is difficult to identify the number of characters in a word. After the locating the character they measure the distance for feature comparison based on height, width, Distance between character with in a word, Distance between words. Based on this distance measurement they find the possibility of number of character in a word and based on this, apply align mode of character. By measuring the distance between character, they find the approximate number of characters in a single word using the maximum no of characters in a single word using the maximum no of character possibility they find the missing character. How many characters can be inserted in the space is calculated by considering distance between two neighbouring characters.

VII. CONCLUSION

In this study, main approaches used in the retrieval of character during the last decade are overviewed. Different pre-processing segmentation techniques and various classifiers with different features are also discussed. It is found that neither the structural nor the statistical information can represent a complex pattern alone. Therefore one need to combine statistical and structural information supported by the semantic information supported by the semantic information HMM are very successful in combining information for many retrieval problem. Based on the probability we can easily find the missing character for the retrieval of character. These retrieval methods can be more effective when we used the probability and the some features of the character to retrieve the missing character.

www.ijmas.com

REFERENCES

i. Mohand Alata and Mohammad Al-Shabi, “Text Detection and Character Recognition Fuzzy Image Processing,” Journal of Electrical Engineering, Vol. 57, No. 5, 2006, 258-267.

ii. Qinru Qiu, Qing Wu, and Richard Linderman, “Proceeding of International Joint Conference on Neural Network, San Jose, California, USA, July 31- August 5, 2011.

iii. Akihito Kitadai, Masaki Nakagawa and Hajime BABA, “Similarity Evaluation and shape Feature Extraction for Character Pattern Retrieval to Support Reading Historical Documents,” IEEE International Workshop on Document Analysis Systems. Apr 2012.

iv. Khurram Khurshid, Claudie Faure and Nicole Vincent, “ Word spotting in historical printed documents using shape and sequence comparisons,”Elsevier Nov-2011.

v. Aman Parkash Singh and Sandeep Kaushal, “Retrieval of Degraded Character in the Document”

vi. Gaurav Kumar and Pradeep Kumar Bhatia.” Neural Network based Approach for Recognition of Text images,” International Journal of Computer Application volume 62-No. 14, Jan 2013.

vii. Dinesh Dileep, “ A Feature Extraction Technique Based on character Geometry For Character Recognition,” .

viii. Sumedha B. Hallale and Geeta D. Salunke, “ Twelve Direction Feature Extraction for Handwritten English Character Recognition”, International Journal of Recent Technology and Engineering volume-2, Issue-2, May 2013.

ix. Sandeep Saha, Nabarag Paul and Sayam Kumar Das, “Optical Character Recognition using 40-point Feature Extraction and Artificial Neural Network,” International Journal of Advanced Research in Computer Science and Software Engineering, volume 3, Issue 4, April 2013.

www.ijmas.com