Text Detection and Extraction from Document Images using K-Nearest Neighbor Rule

Mallikka Rajalingam
School of Computer Sciences
Universiti Sains Malaysia
Pulau Penang, 11800, Malaysia
Email: mallikka2002 {at} yahoo.com

Putra Sumari
School of Computer Sciences
Universiti Sains Malaysia
Pulau Penang, 11800, Malaysia

Valliappan Raman
School of Computer Sciences
Universiti Sains Malaysia
Pulau Penang, 11800, Malaysia

Abstract— **Text extraction and text line detection is the foundation of document image analysis. Since many years, a large number of text detection methods have been proposed, where these methods depend on convinced assumptions of documents with various font style, font size, distorted text, uneven lighting, complex background and low resolution. In this paper, reveals k-nearest neighbor rule as a generic text-line detection and text extraction approach that can be applied on a complex mail document images. The performance evaluation of transition map generation and it compares with other two models is presented in this paper. Experimental analysis shows that image based text Optical Character Recognition (OCR) method is to extract the text from the colorful image and detection of advertised mails is very efficient than that of the other existing methods.**

Keywords- Text extraction, Text-line detection, KNN rule, and mail document image.

I. INTRODUCTION

Owing to the recent progress in internet technology, massive amount of text and image e-mails are generated such as attachments and advertisements which contains either text information, pictures, or both. Spammers are sending unwanted e-mails with image content to deceive anti-spam solution. Image-based spam e-mails are entering without user approval and filling their mailboxes with unsolicited unwanted e-mails. Owing to these, and also due to the broad storage space, the anti-spam filtering technique is growing tremendously.

Image-based spam or Image spam is a kind of e-mail spam where the message text of the spam is presented as a picture in an image file. Since most modern graphical e-mail client software will give the image file by default, presenting the message image directly to the user, it is highly effective to overcome normal e-mail filtering software. Image spam e-mail, also known as junk e-mail, is a group of spam that involves nearly similar messages sent to numerous recipients by spammers.

The first time image spam was released into the world by 2003 and most popular in short span of duration. The developers of spam filters, spammers introduced in 2006 [7] tells the tricks of image spam which helps to remove unwanted information from e-mail's body. The spam message embedded into an image which is sent as attachment to numerous recipients by email. In August 2010, report from "Message Labs Intelligence" calculated 92% of e-mail messages to be spam [8]. Since 15 years before spam e-mail is active whereas from 2006, image spam e-mail is active to fool anti-spam filter. Fig. 1 shows sample picture of image spam e-mail which contains text in image taken from internet.

In order to circumvent spammers, anti-spam techniques for filtering unwanted image e-mails are required. Classification is a process of detecting spam and ham e-mails and filtering from mailboxes.

Figure 1. Sample picture of image based Text (sample picture, 2014) [13]

Recent research on classification of image based Text however, is more inclined towards improves filtering accuracy.

Due to cluttering process, the text is not understandable and takes longer time to solve the puzzle; this is the disadvantage of existing technique. Distorted text can be identified by human rather than computer program. The task of Optical Character Recognition (OCR) is to extract and analyze the text embedded into images. The challenges of OCR are to find distorted text appear in images while being discerning enough to analyze between distorted text from background images.

The paper is organized as follows. Section II reviews some background knowledge of proposed work. Section III describes the text extraction and text line detection. Section IV explains about KNN nearest neighbor algorithm. Section V shows analysis and comparison of spectral pattern recognition and OCR. Finally, section VI concludes the paper.

II. BACKGROUND

At first, the Optical Character Recognition (OCR) is a technique to convert scanned or photography images either electronically or mechanically. OCR technique to expanding telegraphy and creating reading devices for the blind. Later it has been made available in online service. The following thing has been made by the OCR are:

1. Preprocessing which includes binarization, line removal, line and word detection, segmentation and normalization.
2. Character Recognition has been done by matrix matching algorithm like pattern matching and feature extraction algorithms such as k nearest neighbor classifiers.
3. Post processing for increasing the accuracy of detecting text.

Usually the text based image classifier method is based on the color, edge and texture-based features. There is a color transition between the text and the background color.

Table 1. Related work with accuracy

The transition had been finding out by the transition maps. After finding the transitions, the edge based detection algorithm is used to find the edge. Later they extract the text strings based on local threshold value. Table 1 shows related existing works with accuracy.

In occurring at beginning of classification process, text overlapped with in image will be incorrectly classified as image since the touching text and images form a whole large connected component. We need to search further the images for these touching characters. The Fig. 2 shows our sample input image used for experimentation.

Author	Year	Method/ Algorithm	% of accuracy
R. Chandrasekaran, & RM. Chandrasekaran [9]	2011	SVM	93% - Extraction
Aria Pezeshk and Richard L. Tutwiler [10]	2011	Hidden Markov Model	93.93% - Recognition
Hyung Il Koo and Nam Ik Cho [11]	2012	State Estimation	99.52% - Detection
Jerod J. Weinman, Zachary Butler, Dugan Knoll, and Jacqueline Field [12]	2014	Expectation-Maximization Fitting Algorithm	33.71% - Recognition

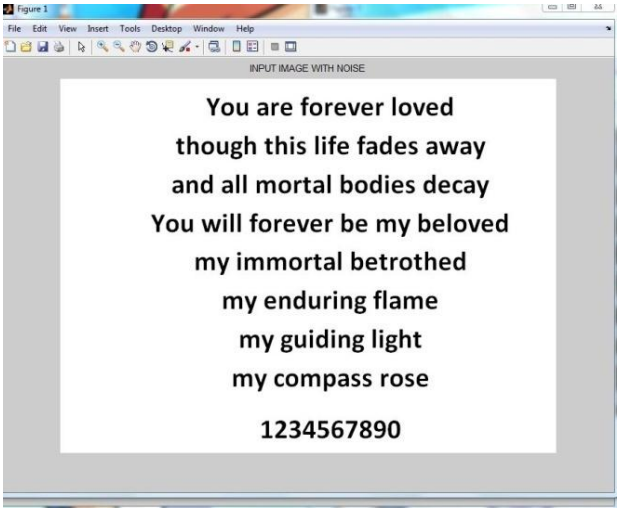

Figure 2. Input image

III. TEXT EXTRACTION

The simplest method for extraction of text from the background is thresholding. Global thresholding [2] techniques extract objects from images having uniform background. Local thresholding methods [5, 3] are suggested to withstand the adverse effect of varying background but at a price of processing cost. The processing cost can be reduced by capturing the regions containing text and then thresholding only those regions instead of thresholding the entire document image. Text-regions in a document image can be detected either by connected component analysis or by texture analysis method. Texture based methods detect the text regions based on the fact that text and background have different textures [10]. In general texture based methods are more robust than the connected component based methods in detecting text regions in documents with complex background but the computational complexity is high. Developers [6] proposed a method to detect the vertical and horizontal edges in an image. They used different dilation operators for the two kinds of edges. Real text regions are then identified using support vector machine. Researchers [4] extracted text as those connected components that follow certain size constraints. Automatic text information extraction based on key words or field of interest. Experiment result showed that proposed method perform better than the random and the initial lead method [14]. As per the thresholding value the text has been converted in the Fig. 3. Another method of text extraction is Transition Map Generation.

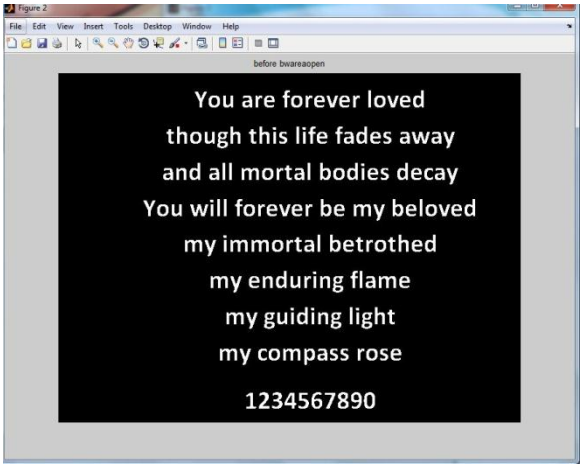

Figure 3. Extraction of Text

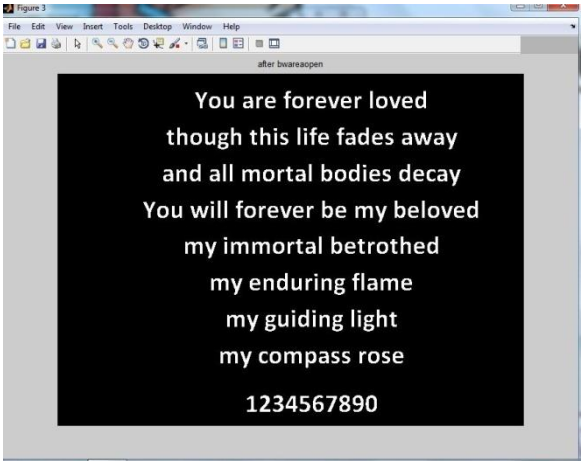

Figure 4. Connected components reshaped

A. Transition Map Generation

The intensities of three consecutive pixels are decreasing logarithmically at the boundary of bright overlay text due to color bleeding. It is also observed that the intensities of three consecutive pixels increases exponentially at the boundary of dark overlay text. Graphical illustration of intensity change in the transition region is also shown in Fig. 7. We first containing "Bright to Dark (B-D)" and "Dark to Bright (D-B)" transitions, respectively. 20 transition areas sampled from each image is averaged and illustrated in the figure. Since the change of intensity at the boundary of overlay text may be small in the low contrast image, to effectively determine whether a pixel is within a transition region, the modified saturation is first introduced as a weight value based on the fact that overlay text is in the form of overlay graphics. The modified saturation is defined as follows:

$$MS(x,y)=1-\frac{3}{Re+Gr+Bl[min(Re,Gr,Bl]} \quad (1)$$

Where *MS(x,y)* is the saturation value. The transition map can be utilized as a useful indicator for the overlay text region. To generate the connected components, we first generate a linked map (1). If a gap of consecutive pixels between two nonzero points in the same row is shorter than 5% of the image width, they are filled with 1s. If the connected components are smaller than the threshold value, they are removed. The threshold value is empirically selected by observing the minimum size of overlay text region. Then each connected component is reshaped to have smooth boundaries as shown in the Fig. 4.

B. Text Line Detection

The previous stage has a high detection rate but relatively low precision due to many false positives. This means that most of the text lines are included in the text area boxes while at the same time some bounding boxes may include more than one text line as well as noise. The noise usually originates from objects with high intensity edges that connect to the text lines during the dilation process. The low precision also originates from detected bounding boxes which do not contain text but objects with high vertical edge density. To improve precision by rejecting the false alarms we use a method based on horizontal and vertical projection. Fig. 5 shows taxonomy of text and image based emails.

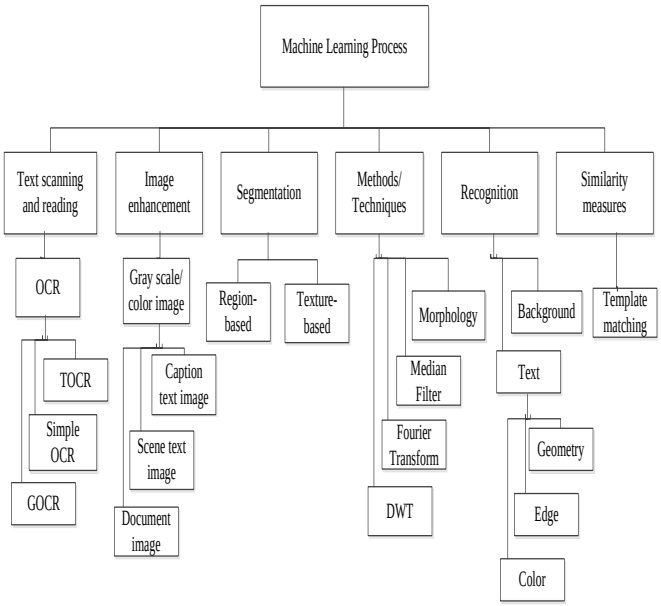

Figure 5. Taxonomy of text and image based emails

Firstly, the horizontal edge projection of every box is computed and lines with projection values below a threshold are discarded. In this way, boxes with more than one text line

are split and some lines with noise are also discarded as shown in Fig. 7. Nevertheless, boxes which do not contain text are usually split in a number of boxes with very small height and discarded by the next stage due to size constraints.

IV. NEAREST NEIGHBOR

The ***k*-Nearest Neighbors algorithm** (or ***k*-NN** for short) is a non-parametric method used for classification and regression. In both cases, the input consists of the *k* closest training examples in the feature space. The output depends on whether *k*-NN is used for classification or regression:

- In *k-NN classification*, the output is a class membership. An object is classified by a majority vote of its neighbors, with the object being assigned to the class most common among its *k* nearest neighbors (*k* is a positive integer, typically small). If $k = 1$, then the object is simply assigned to the class of that single nearest neighbor.
- In *k-NN regression*, the output is the property value for the object. This value is the average of the values of its *k* nearest neighbors.

k-NN is a type of instance-based learning, or lazy learning, where the function is only approximated locally and all computation is deferred until classification. The *k*-NN algorithm is among the simplest of all machine learning algorithms.

KNN for Classification

Let's see how to use KNN for classification. In this case, we are given some data points for training and also a new unlabelled data for testing. Our aim is to find the class label for the new point. The algorithm has different behavior based on k.

Case 1 : k = 1 or Nearest Neighbor Rule

This is the simplest scenario. Let x be the point to be labeled. Find the point closest to x; Let it be y. Now nearest neighbor rule asks to assign the label of y to x. This seems too simplistic and sometimes even counter intuitive. If you feel that this procedure will result a huge error, you are right but there is a catch. This reasoning holds only when the number of data points is not very large.

If the number of data points is very large, then there is a very high chance that label of x and y is same. An example might help – Let's say you have a (potentially) biased coin. You toss it for 1 million time and you have got head 900,000 times. Then most likely your next call will be head. We can use a similar argument here.

Let's try an informal argument here, assume all points are in a D dimensional plane. The number of points is reasonably large. This means that the density of the plane at any point is fairly high. In other words, within any subspace there is adequate number of points. Consider a point x in the subspace which also has a lot of neighbors. Now let y be the nearest neighbor. If x and y are sufficiently close, then we can assume that probability that x and y belong to same class is fairly same. Then by decision theory, x and y have the same class. One of their striking results is to obtain a fairly tight error bound to the Nearest Neighbor rule. The bound

$$P^* \leq P \leq P^*(2 - \frac{c}{c-1}P^*)$$

Where $P^*$is the Bayes error rate, c is the number of classes and P is the error rate of Nearest Neighbor. The result is indeed very striking because it says that if the number of points is fairly large then the error rate of Nearest Neighbor is less than twice the Bayes error rate.

Case 2 : k = K or k-Nearest Neighbor Rule

This is a straightforward extension of 1NN. Basically what we do is that we try to find the k nearest neighbor and do a majority voting. Typically k is odd when the number of classes is 2. Let's say k = 5 and there are 3 instances of C1 and 2 instances of C2. In this case, KNN says that new point has labeled as C1 as it forms the majority. We follow a similar argument when there are multiple classes.

One of the straight forward extensions is not to give 1 vote to all the neighbors. A very common thing to do is *weighted kNN* where each point has a weight which is typically calculated using its distance. For example, under inverse distance weighting, each point has a weight equal to the inverse of its distance to the point to be classified. This means that neighboring points have a higher vote than the farther points. It is quite obvious that the accuracy might increase when you increase k but the computation cost also increases.

V. ANALYSIS

The text images and background images are measured with the use of the intensity value of them. We can feel the intensity of the images by a histogram. By our transition mapping we have find out the intensities value varied between our text and the background. The histogram value of the sample image is shown in Fig. 6.

The graphical representation is very easiest way to compare and distinguish the characteristics of two different models. Spectral pattern matching is the existing digital image classification method which is used to classify the images, separate the images and texts are used in the various applications. Our proposed Image based Text OCR recognition is the method to extract the text from the colorful

image and detection of advertised mails is very efficient than that of the other existing methods.

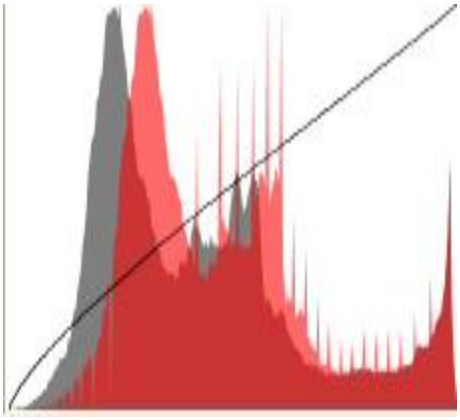

Figure 6. Histogram picture of our sample

The spectral pattern matching is extremely high dimensionality in factorial representation of spectral data. Dependency of the dimensions, where dependency may be among adjacent dimensions or those far apart. Lack of credible reliability measurement associated with recognition decisions. So the comparison of OCR and Spectral pattern recognition is shown in the Fig. 7.

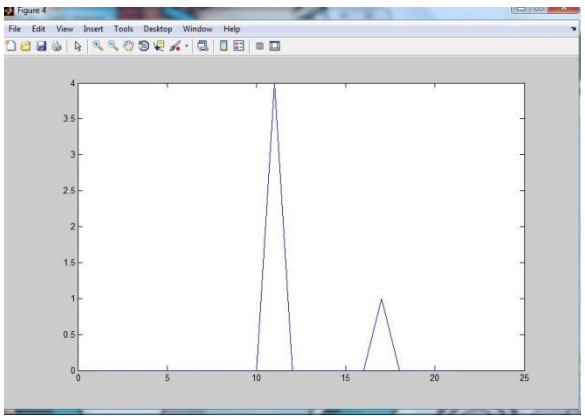

Figure 7. Comparison of Spectral pattern recognition and OCR.

The performance of each method is an indication of the usefulness of the method's underlying assumptions. For example, the global fixed method assumes that the scanned gray levels are very well normalized. The Sauvola–Niblack method uses parameters *k* and *R* that may not be optimal for the statistical properties of historical documents. The Chang method assumes that the ideal threshold is exactly midway between the lightest and darkest pixels in a block. The methods based on the Otsu algorithm do best. The main assumption behind Otsu approaches is that there are two classes of pixels, which is always true globally, leaving only problems with local illumination. The multi-resolution variant (MROtsu) adapts to illumination changes by making local decisions, using the assumption of a 2:1 white/black ratio to establish the smallest valid scale where two classes are present. This assumption is valid for this data set, and should hold for most printed text. The error diffusion method was competitive, but its error passing did not improve upon the global Otsu threshold upon which it was based. There are several possible contributions to this:

- Purely local algorithms cause undesired results in background areas, affecting text-detection capabilities of the OCR black box
- The simplification of using blocks instead of a sliding window may reduce the ability to adapt the sharp illuminance changes (this also affects the MROtsu method and the implementation of the Chang method)
- Many images have poor resolution, quantization, or are noisy
- In general, the Sauvola–Niblack images were too light, causing more broken characters

VI. CONCLUSION

A method of detecting and extracting text connected to images has been proposed in this paper. The proposed method merges with text line detection to interpret connection of text and images. Experimental result shows that the proposed method improved the percentage of correctly extracted text from images as well as the accuracy of OCR significantly. The characters can also be in different style, orientation, font size, complex background, uneven lighting, distorted text and low resolution. In the near future, hope to develop existing method to Support Vector Machine (SVM) algorithm with OCR. The integration of classification and text extraction may be one promising approach to separate the overlapped characters from images.

ACKNOWLEDGMENT

Sincere thank and recognition goes to my advisor, Associate Professor, Dr. Putra Sumari, who guided me through this research, inspired and motivated me. Thanks for RCMO, USM to continuously provide support for conducting research. We also thank the Universiti Sains Malaysia (USM) for supporting this research.

REFERENCES

[1] Wonjun Kim and Changick Kim. A New approach for overlay text detection and extraction from complex video scene. IEEE Transactions on Image Processing, 18(2). 2009.

[2] N Otsu. A threshold selection method from gray level histograms. *IEEE Transaction Systems, Man & Cybernetics, 9(1):62-66, 1979.*

[3] J. Sauvola and M. Pietikainen. Adaptive document image binarization. *Pattern Recognition, vol.*33, 225-236, 2000.

[4] Y. Zhong, K. Karu, and A. K. Jain. Locating text in complex color images. *Pattern Recognition*, 28 (10): 1523–1536, 1995.

[5] W. Niblack. An introduction to image processing. Printice Hall, Englewood Cliffs, NJ. pp. 115-116, 1986.

[6] C. Datong, H. Bourland, and J. P. Thiran. Text identification in complex background using SVM. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol. 2, pp. 621–626, 2001.

[7] B. Biggio, G. Fumera, I. Pillai, and F. Roli. Image spam filtering by content obscuring detection. Fourth Conference on Email and Anti-Spam (CEAS 2007). August 2-3. Mountain View California USA, 2007.

[8] D. DeBarr and H. Wechsler. Spam detection using random boost. ELSEVIER. Pattern Recognition Letter. Vol.33. pp.1237-1244, 2012.

[9] R. Chandrasekaran and RM. Chandrasekaran. Morphology based text extraction in images. International Journal of Computer Science and Technology. Vol.2, 2011.

[10] Aria Pezeshk and Richard L. Tutwiler. Automatic feature extraction and text recognition from scanned topographic maps. IEEE Transactions on Geoscience and Remote Sensing. 49(12), 2011.

[11] Hyung Il Koo and Nam Ik Cho. Text-Line extraction in handwritten chinese documents based on an energy minimization framework. IEEE Transactions on Image Processing. 21(3), 2012.

[12] Jerod J. Weinman, Zachary Butler, Dugan Knoll, and Jacqueline Field. Toward integrated scene text reading. IEEE Transactions on Pattern Analysis and Machine Intelligence. 36(2), 2014.

[13] Sample picture. https://www.google.com/textandimageemails_CCU. 23 April 2014.

[14] Mahmoud B. Rokaya. Automatic text extraction based on field association terms and power links. International Journal of Computer and Information Technology (IJCIT). 2(6): 1039-1053, 2013.