ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 5, Issue 2, February 2018

Unconstrained Handwritten Document Retrieval Based on User Query Interaction

[1] Dr.V.C.Bharathi, [2] Dr. K. Vaidhei
[1][2] Associate Professor
[1] CSE, Madanapalle Institute of Technology and Science, Madanapalle, Andra Pradash, India.
[2] CSE, Stanely College of Engineering and Technology for Women, Hyderbad, Teluguna, India.

Abstract - **In unconstrained handwritten document retrieval given a list of documents, retrieve the documents based on user query keyword and find the similar keyword in the relevant document that can be search and retrieved handwritten documents with efficient information. The work involves preprocessing of the input document and segmentation is applied to the document based on contour to segment the individual words. In relevant index stores all information of the words, it contains relevant information of the document, the position of the words and class label of each word. In this paper, we proposed unconstrained document retrieval based on user query. After indexing the segmented word images partitioned into 2×2 subblock, each subblock region again partitions into 5×5 subblock. In each subblock, to calculate average intensity of pixels and to find the maximum average values in horizontal and vertical direction. Thereby 40-dimensional features are extracted from 2×2 subblock and extracted features are fed to SVM with RBF kernel to construct the models for all classes. In testing samples, a user is given the query in the search area. The user query keyword randomly selected the corresponding word image in testing samples and to extract the feature for the word. The extracted features are fed for testing to retrieve the appropriate class. The class label is used to retrieve the corresponding index information and retrieve the information from the list of document.**

Keywords: Handwritten Word Retrieval; Word Spotting; Segmentation; Maximum Intensity Vector (MIV); Support Vector Machine.

I. INTRODUCTION

In the modern world, large volume of document images is easily available and the popular use of the web demands the technology for efficient document retrieval based on keywords. Text search relies on the keyword identification, and so, it has been restricted to the printed documents. The problem of word spotting in handwritten documents has attracted a lot of attention in the community in the last few years. It consists of detecting any given keyword in document images. This task is important in numerous applications, such as querying textual handwritten documents, automatic document categorization, indexation, information retrieval in handwritten document databases. Handwritten keyword identification is the pattern classification task which identifies the keyword presented in the handwritten documents. Keyword identification is a major task of indexing and retrieval methods, it can be defined as searching and locating given query information in a large collection of significant information. In document analysis, the research area is mainly focused on assigning index of all words, thereby searching and retrieving the documents fastly.

II. RELATED WORK

Simon Thomas et al.[1] proposed query based multiple keyword spotting in handwritten documents using HMM. In this method, each text line images is divided into vertical overlapping windows frames taken as input directly from the frames. GMM-HMM and MLP-HMM based system, a feature vector is computed for each frame to construct the models in HMM and determined to maximize the word recognition rate on a validate of the testing sample. Raid Saabni et al.[2] is presented for spotting and retrieval keywords in handwritten documents based on similarities between the words using dynamic time algorithm to spotted the given word.

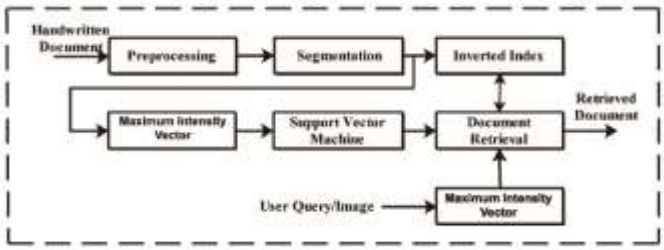

Fig 1. Unconstrained Handwritten Document Retrieval

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 5, Issue 2, February 2018

Ismet Zeki Yalniz [3] is proposed to search text in scanned books. The given query word image aim is to retrieve matching words in the books based on similarity of the SIFT descriptors are extracted over the corner points of each word image consider as features. In hierarchical K means algorithm used to identified the cluster and index location from inverted file for query resolution. Chew Lim Tan et al.[4] is presented image based retrieval and keyword spotting in handwritten documents. Based on local pixel direction contribution is extracted from each cell grid and character shapes coding to extract the features are spotted and retrieve the documents using HMM and TF/IDF similarity. Volmar Frinken [5] proposed retrieving all instances of a given keyword from a document based on modification of the CTC token passing algorithm with recurrent neural network. The systems outperform a modern keyword spotting system based on HMM and DTW approach. Jon Almazan et al.[6] described an unsupervised segmentation free method for word spotting in document Images. Documents are represented with a grid of HOG descriptors and sliding window approach is used to locate the document region that are most similar to the query and use Exemplar SVM produce better representation of the query in an unsupervised way. Davif Fernandez et al.[7]proposed handwritten word spotting in old manuscript images. In this method, after segmentation the word images features are extracted from region of the shape of the contour and blurred shape model feature in second level is to extract based on pixel distribution of the words with similar pixel distribution in various cluster.

III. OUTLINE OF WORK

This paper deals with unconstrained handwritten document retrieval based on user query interaction. The method has been evaluated with handwritten data set collected from different writer with various writing styles. Users interact with document based on queries. The approach tries to match the given query word and similar word in the relevant document search and retrieve the unconstrained document. The rest of the paper is organized as follows; section 2 describes proposed unconstrained handwritten document retrieval. In this approach preprocessing and segmentation task involved. Maximum Intensity Vector is describes in section 3. Section 4 describes inverted index, Section 5 deals with Support Vector Machine. Section 6 shows the experimental results of our approach and section 7 conclusion of the work.

IV. PROPOSED UNCONSTRAINED HANDWRITTEN DOCUMENT RETRIEVAL

The given input document is preprocessed with binarization and morphological filtering to eliminate the noise from unwanted effects of the original image. The original image is dilated for word segmentation based on the contour to segment the individual words for feature extraction. Segmented word image are subblock into 2x2 subblock and each subblock partition into different size of subblock. In each subblock calculate average intensity of pixels and find maximum average intensity value in horizontal and vertical direction to extract feature and inverted index can be assign for all words, it contains the document information, position of the word and class label. Fig. 1 shows a block diagram of the query based unconstrained handwritten document retrieval proposed in this paper. This block diagram consists of the following steps:

A. Input Document

Handwritten dataset created by different writers are obtained, to ensure various writing styles across different age group and different genders. The input images are captured using a scanner with 300 dpi resolution and saved in JPEG/JPG format as shown in Fig. 2. Handwritten words are collected from 80 different writers with different document, writing styles and size as shown in Table 1.

TABLE I HANDWRITTEN DOCUMENT

S.No	Gender	Age	Number of Samples
1	Male	25-40	35
2	Female	18-26	35
3	Children	9-14	10

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 5, Issue 2, February 2018

B. Preprocessing

The role of the preprocessing is required for removing unwanted noises from the input documents. The noise is introduced into the image while the acquisition of images. Preprocessing is to reduce noise and to improve the quality of the handwritten document for robust recognition and to ensure the handwritten text is in a suitable form. Binarization and morphological filtering are applied in preprocessing stage.

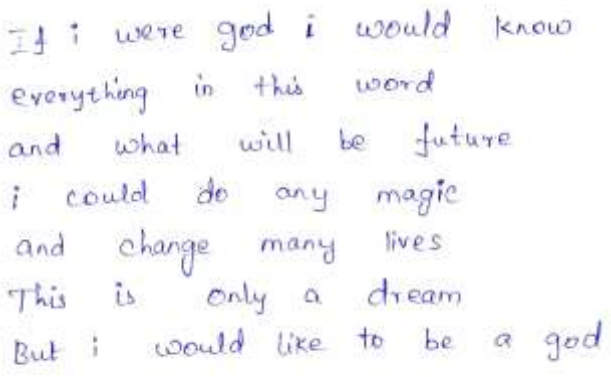

Fig 2. Input Document

Binarization

The original RGB image is converted into gray scale and then the gray image is converted into binary image. For word segmentation, binary image is complimented as white foreground and black background.

Morphological Filtering and Dilation

Input document taken in a different environment that some of them might contain noises and shadows. Morphological filtering is applied over the binary image to remove the small objects from the binary image. Binary area open method is a type of morphological filtering which removes the pixels objects that accommodate fewer than 30 pixels [8]. The filtered images are applied for dilation, to enlarge foreground objects and diminish background for segmentation of the word.

Segmentation

Segmentation is the process of extracting isolated word from the binary image [9]. In dilated image to find the connected components of the foreground object and measure the properties of labelled region by placing the bounding box over the binary image as shown in Fig. 3. Segmenting the bounding box contours which are various size for feature extraction uniformly re-sized to 120 width x 120 height.

V. MAXIMUM INTENSITY VECTOR

The segmented words are partitioned into 2 x 2 subblock each of size 60 rows and 60 columns block size. In each 60 rows and 60 columns region is subdivide into 5 x 5 subblock each of 12 rows and 12 columns. In each subblock calculate average intensity of pixels and find the maximum average intensity value in horizontal and vertical direction as shown in Fig 4(a) and Figure 4(b) of each 5 x 5 subblock. There by 40 dimension features are extracted from 2 x 2 subblock.

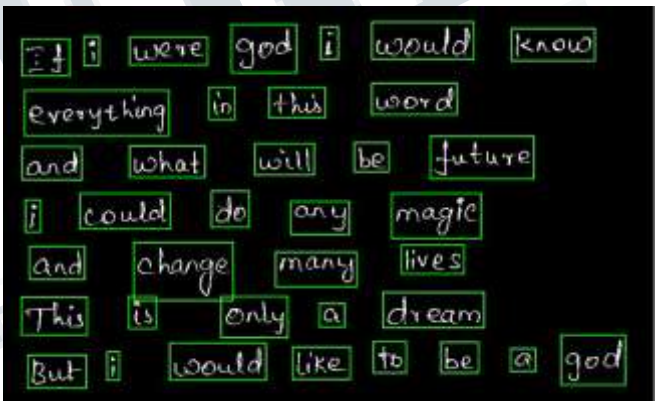

Fig 3. Word Segmentation

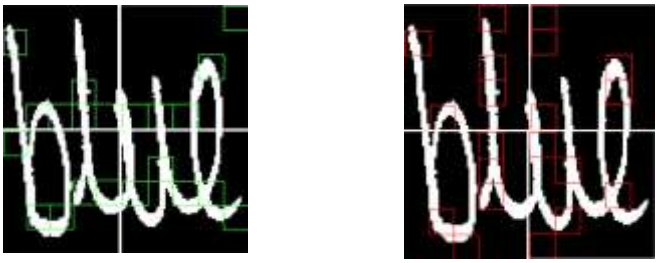

Fig 4. a) MIV in Horizontal b) MIV in Vertical direction

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 5, Issue 2, February 2018

VI. INVERTED INDEX

The inverted index architecture consists of a lexicon, which stock all the indexed terms in the collection of information about this word. The lexicon accommodates the location of each word in posting list, which holds the position of each occurrence of that word occurs in that document as shown in Fig 5. The postings list of the document numbers will be processed sequentially from the beginning of the file, the list can be stores as the starting position followed by a list in document identifiers [10]. For example, the user given phrase query string 'i', for each posting store the positions within each document where the word occurs. For each word postings list is [Document No, Position]. The given query string 'i' occurs four times in a document and posting information [1:2, 1:5, 1:17, 1:32] to retrieve the document position for all unconstrained handwritten documents.

VII. SUPPORT VECTOR MACHINE

Support Vector Machine (SVM) is a group of supervised learning methods that can be applied to classification and regression. It is a popular technique for classification in visual pattern recognition [11, 12]. The SVM is mostly used in kernel learning algorithm. It achieves reasonably vital pattern recognition performance in optimization theory [13, 14]. A classification task is typically involved with training and testing data. The training data are separated by (x1, y1),

(x2, y2),…..,(xm,ym) into two classes, where xi ∈Rn contains

n-dimensional feature vector and yi ∈ {+1,-1} are the class labels. The aim of SVM is to generate a model which predicts the target value from testing set. In binary classification the hyper plane w.x + b = 0 where w ∈Rn, b ∈Rn is used to separate the two classes in some space Z. The maximum margin is given by M=2/|w|.

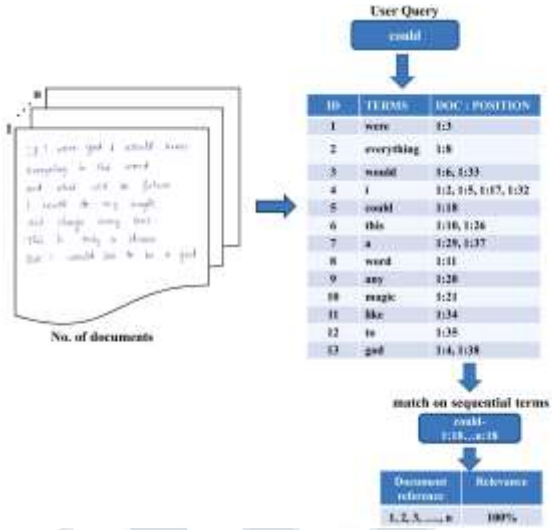

Fig. 5.Inverted Index

VIII. EXPERIMENTAL RESULTS

This section describes the experiments conducted in maximum intensity vector features that use SVM classifier for document retrieval.

A. Handwritten Dataset

In this experimental work, the input handwritten document sample images are collected from 80 different writers with different writing styles and size. The handwritten documents were scanned at dimensions of pixels, 300 dpi images. In the proposed work, 57 writers sample documents were taken for training and 23 writers sample documents were taken for testing. The training and testing samples are used to measure the performance of maximum intensity vector.

B. Evaluation metrics

The retrieval system evaluated based on the recognition accuracy defined as

$$\text{Recognition Accuracy (RA)} = N_R/N_T \quad (1)$$

$$\text{Error Rate (ER)} = N_M/N_T \quad (2)$$

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 5, Issue 2, February 2018

Where NR denotes the number of query phrase word for which relevant documents were retrieved, NM denotes the number of query phrase misclassified in relevant documents and NT denotes the total number of query phrase word in relevant document.

When user given the query phrase, phrase are randomly selected from the testing samples to consider the corresponding word image as the search keyword. The word image to extract the features using maximum intensity vector, the features are fed to SVM with RBF kernel to identify the corresponding class as assign as index id and to retrieve the position of each occurrence of that word occurs in posting list of the inverted index. The retrieval position are spotted in a document and retrieve the relevant document.

The 40 dimension MIV features are fed to SVM with RBF kernel to identify the corresponding class, each class contains a label of the word and spot the corresponding label position of each occurrence of that word in positioning list. To retrieve the position is spotted and to count the occurrences of relevant documents.
Recognition accuracy (RA) and Error rate (ER) of the user query phrase are shown in Table 2. In the unconstrained handwritten document, user query phrase to predict number of times it appears in each document is evaluated based on Eq. 6. It considers single occurrence word 'change', two occurrence word 'this' and four occurrence word 'i' to measure the performance of predicted accuracy as shown in Fig 6.

TABLE III RECOGNITION ACCURACY AND ERROR RATE OF USER QUERY.

Query Phrase	RA(%)	ER(%)
dream	95.65	4.34
word	82.60	17.39
could	86.95	13.04
magic	91.30	8.69
god	86.95	13.04
lives	78.26	21.73
but	86.95	13.04
were	91.30	8.69
this	86.95	13.04
everything	100	0
know	78.26	21.73
could	86.95	13.04
change	95.65	4.34
future	86.95	13.04
Average	**88.19**	**11.80**

IX. CONCLUSION

In this paper the proposed a novel approach for unconstrained handwritten document retrieval based on user query interaction. The method has been predict the user query based document retrieval and number of times user query occurrences in each document. The methods has been implemented the following steps that include segmentation, maximum intensity vector in horizontal and vertical direction, all the segmented word assign the index in inverted for position of the words stored and SVM with RBF kernel are trained and tested the user query to retrieve the document. The performance of the unconstrained handwritten document retrieval, average accuracy obtained based on user query 88.19%, also based on number of character in word and numbers of occurrences in document experimental work is done in this proposed method. In future work, intended to enhance for document retrieval based on multi query user interaction.

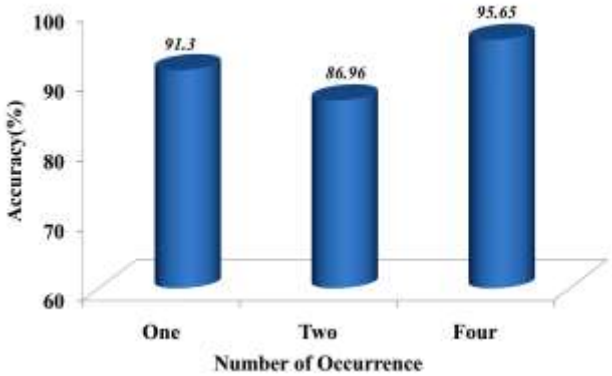

Fig.6. Predicted accuracy of the word occurrences

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 5, Issue 2, February 2018

X. CONCLUSION

In this paper the proposed a novel approach for unconstrained handwritten document retrieval based on user query interaction. The method has been predict the user query based document retrieval and number of times user query occurrences in each document. The methods has been implemented the following steps that include segmentation, maximum intensity vector in horizontal and vertical direction, all the segmented word assign the index in inverted for position of the words stored and SVM with RBF kernel are trained and tested the user query to retrieve the document. The performance of the unconstrained handwritten document retrieval, average accuracy obtained based on user query 88.19%, also based on number of character in word and numbers of occurrences in document experimental work is done in this proposed method. In future work, intended to enhance for document retrieval based on multi query user interaction.

XI. REFERENCES

[1] Thomas, Simon, Clement Chatelain, Laurent Heutte, Thierry Paquet, and Yousri Kessentini. "A Deep HMM model for multiple keywords spotting in handwritten documents", Pattern Analysis and Applications, pp. 1-13, 2012.

[2] Saabni, Raid M., and Jihad A. El-Sana. "Word spotting for handwritten documents using Chamfer distance and dynamic time warping." In IS T/SPIE Electronic Imaging, pp. 78740J-78740J. International Society for Optics and Photonics, 2011.

[3] Yalniz, Ismet Zeki, and Raghavan Manmatha. "An efficient framework for searching text in noisy document images." In Document Analysis Systems (DAS), 2012 10th IAPR International Workshop on, pp. 48-52. IEEE, 2012.

[4] Tan, Chew Lim, Xi Zhang, and Linlin Li. "Image Based Retrieval and Keyword Spotting in Documents." In Handbook of Document Image Processing and Recognition, pp. 805-842. Springer London, 2014.

[5] Frinken, Volkmar, Andreas Fischer, R. Manmatha, and Horst Bunke. "A novel word spotting method based on recurrent neural networks." Pattern Analysis and Machine Intelligence, IEEE Transactions on 34, no. 2 (2012):

[6] Almazn, Jon, Albert Gordo, Alicia Forns, and Ernest Valveny. "Segmentation-free word spotting with exemplar SVMs." Pattern Recognition 47, no. 12 (2014): 3967-3978.

[7] Fernndez, David, Josep Llads, and Alicia Forns. "Handwritten word spotting in old manuscript images using a pseudo-structural descriptor organized in a hash structure." In Pattern Recognition and Image Analysis, pp. 628- 635. Springer Berlin Heidelberg, 2011.

[8] Bharathi, V. C., and M. Kalaiselvi Geetha. "Hierarchical Character Grouping and Recognition of Character Using Character Intensity Code." In Artificial Intelligence and Evolutionary Algorithms in Engineering Systems, pp. 789-797. Springer India, 2015.

[9] Bharathi, V. C., and M. Kalaiselvi Geetha. "Performance Evaluation of GMM and SVM for Recognition of Hierarchical Clustering Character." In Advanced Computing, Networking and Informatics-Volume 1, pp. 161-169. Springer International Publishing, 2014.

[10] Ferguson, Paul, and Alan F. Smeaton. "Index ordering by query-independent measures." Information Processing and Management 48, no. 3 (2012): 569-586.

[11] N.Cristianini, J.Shawe-Taylor , An Introduction to Support Vector Machines and Other Kernel-based Learning Methods Cambridge University Press, (2000).

[12] T.Mitchell, Machine Learning, McGraw-Hill Computer science series, (1997).

[13] V.Vapnik, Statistical Learning Theory, Wiley, NY, (1998).

[14] J.P Lewis, Tutorial on SVM, CGIT Lab, USC, (2004).