An Efficient Approach for Text Query Searching and Word Spotting in Low Quality Manuscripts

HARITHA V R
Dept. Of Computer Science and Engineering
M.E.A. Engineering College
Perinthalmanna, Kerala

SREERAM S
HOD
Dept. Of Computer Science and Engineering
M.E.A Engineering College
Perinthalmanna, Kerala

***Abstract*— The standard technique for indexing documents is to scan them in, convert them to machine readable form (ASCII) using Optical Character Recognition (OCR) and then index them using a text retrieval engine. However, OCR does not work well on handwriting. Here, an alternative scheme is proposed for indexing such texts. The key idea behind word spotting is to find all the occurrences of a given word in the document images. Here the input is a word which is usually provided as user input, and the output are all the coordinates of the corresponding word images. Most of the existing approaches to word spotting choose a certain number of word images containing the user query as templates and find the best matches in the dataset. Here for any query, there are a few word images stored as templates in a training set. However, since it is not easy to get all the word images of possible user queries in advance, this assumption limits the application of word spotting to a small set of keywords. To solve this problem, in the approach proposed in this paper, matching is performed in character image level rather than in word image level. The proposed approach provides efficient way of searching text like queries than the existing system.**

Index Terms- Keyword spotting, Offline handwriting

I. INTRODUCTION

Despite the growing use of electronic documents in our daily life, the use of paper documents is still playing an important role. Current technologies allow us convenient and inexpensive means to capture, store, compress and transfer digitilized images of documents. There are lots of historical handwritten documents with information that can be used for several studies and projects. The conversion of historical document collections to digital archives is of prime importance to society both in terms of information accessibility, and long- term preservation. Handwritten documents are used to be found in historical archives. Examples are unique manuscripts written by well known scientists, artists or writers; letters, trade forms or administrative documents kept by parish or municipalities that help to reconstruct historical sequences in a given place or time, etc.

Nowadays thousand of digitized documents are unutilized because they are not indexed. There are some levels of indexation in terms of meta-data, from the naming of the author and the brief history of the book to a full text transcription. Nevertheless, there is not a unique technique that allows us to index the document correctly. During the last decades these techniques have experienced great improvements and the error rates have dropped to a level that makes commercial applications feasible. Traditional optical character recognition (OCR) systems fail to process hand-written documents, and they are only suitable for modern printed documents. However, the off-line handwritten text recognition systems, which take an image of a piece of handwriting as input, are working properly in restricted vocabularies.

The Document Image Analysis and Recognition community is interested in preserving these documents and extracting all the valuable information from them. There are two ways to extract the information: transcribing documents (word-to-word) and word-spotting .The latter approach is well suitable for noisy images. Handwritten word-spotting[1] refers to the problem of detecting specific keywords in handwritten document images. A model is provided as a query, and the goal is to retrieve all the occurrences in a word image database (or regions of a document collection). But, one of the problems of these documents is the access to them. The majority of material is only physically accessible, and only a few of authorized people can access to them.

In the above approaches assume that for any query, there are a few word images stored as templates in a training set. However, since it is not easy to get all the word images of possible user queries in advance, this assumption limits the application of word spotting to a small set of keywords[15]. To solve this problem, in the approach proposed in this paper, matching is performed in character image level rather than in word image level. Since the possible number of characters is very limited, i.e., 26 letters for English text, it will be much easier to build a training set of character images.

In this the input text is matched with template character then the query image is created from template characters. Word images forms with combination of character images stored in the database. The proposed approach provides an efficient way of searching text like queries in document images.

II. RELATED WORKS

There are two types of word-spotting approaches, depending on how the input is specified: Image based methods, also known as "query-by-example", operate through the image representation of the keywords [2], [3], [4], [5], [6]. The recognition based, or "query-by-string" methods, operate with the ASCII representation of the keywords [7], [8], [9], [10], [11], [12]. In the first kind of approaches, the input image is represented as a sequence of features and is matched to a set of template keyword images. The performance of this kind of approaches is limited when dealing with a wide variety of unknown writers. On the contrary, recognition based approaches are not limited to a single writer, at the expense of a more complex matching process, derived from conventional handwriting recognition systems. In this context, many works has focused on several variants of Hidden Markov Models (HMMs) to address this intrinsically sequential problem [12], [10], [11], [8].

In the proposed system, input is in the form of "query by string" methods. But the proposed solution completely avoids machine recognition of handwritten words as this is a difficult task. Our aim is to locate particular word in a handwritten document image.

III. PROPOSED SYSTEM

The idea of word spotting is to find all the occurrences of a given word in the document images. Here the input is a word which is usually provided as user input, and the output are all the coordinates of the corresponding word images. Most of the existing approaches to word spotting choose a certain number of word images containing the user query as templates and find the best matches in the dataset. Thus the most important step is to determine the similarity between two word images. In the above approaches assume that for any query, there are a few word images stored as templates in a training set. However, since it is not easy to get all the word images of possible user queries in advance, this assumption limits the application of word spotting to a small set of keywords. To solve this problem, in the approach proposed in this paper, matching is performed in character image level rather than in word image level. Since the possible number of characters is very limited, i.e., 26 letters for English text, it will be much easier to build a training set of character images.

A.Outline of Algorithm

Step1 : Data acquisition
Step 2 : Pre-processing
Step 3 : Word segmentation
Step 4**:** The extracted words matched against combination of character templates in the dataset.
The matching is done in two phases. First, the number of words to be matched is pruned[13]using the areas and aspect ratios of the word images - the word to be matched cannot have an area or aspect ratio which is too different from the template. Next, the actual matching is done by using a matching algorithm.

Step 5: The matching algorithm rank the matched words in the correct order.
Step 6: Locate the matched words in the particular location.

1)Data Acquisition
This is the stage in which data are collected as part of the word spotting process. The data may be captured online while the user is writing on a digitizer or PDA. In the offline case, the data is obtained by scanning the image after the writing process is over.Fig.1(a)shows an example.

2)Pre-processing
Pre-processing is an important step in word spotting process because of the variations in the writing style among different users and the existence of huge amount of noise in the images after scanning. The pre-processing[14]steps involved in the offline case are:

a)Binarization
Binarization[13]is the process of converting gray scale images to binary images. It is done in order to identify the objects of interest from the image. It separates the foreground pixels from the background pixels.See Fig.1(b)

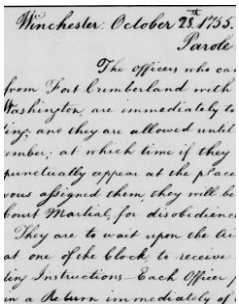
Fig.1(a)scanned document


Fig.1(b)After binarization

b)Noise Removal
A large amount of noise may occur in the image obtained after scanning. This may be due to the poor quality of the scanner or the use of degraded documents. Gaussian noise & Salt and Pepper noise are two such common noises. Filtering techniques such as linear filtering, median filtering or adaptive filtering can be used to remove noise to a certain extent.

c)Normalization
Normalization is the process of converting the image into a standard size[14]. Bilinear and Bicubic interpolation techniques can be used for size normalization.

3)Word Segmentation
A simple technique is used for segmenting words.The technique assumes that a binary image of each page is available and further assumes that the words are white against a dark background (if it is otherwise in the original image, the image can be inverted). Since the spacing between adjacent

characters in a word is smaller than the spacing between adjacent words, a new image is constructed using a smoothing and thresholding operation. If two white pixels are separated by less than a certain distance k, the intermediate pixels are made white. This is done in the horizontal direction khoriz . In the case of handwriting, this procedure also needs to be performed in the diagonal direction - mainly to prevent descenders from breaking up. kdiag . Note that each of these window operations may be viewed as a smoothing and thresholding operation or as a morphological closureo peration. Connected components are now recovered from this image. A minimum bounding rectangle is now constructed using the connected components. The minimum bounding rectangles essentially give a segmentation of the page into words. A number of algorithms exist in the literature for segmenting words from binary images and essentially[13].

B.Using Euclidean Distance Mapping for Matching

There are three steps in the matching:
1. Alignstep: First the images are roughly aligned. In the vertical direction, this is done by aligning the baselines of the two images. The baseline is computed as follows. The difference in the number of white pixels between adjacent scan lines is computed. The point at which the difference is maximum is declared to be the baseline. The baseline computation is performed for both images, and the images then shifted so that they are aligned. In the horizontal direction, the images are aligned by making their left hand sides coincide. The alignment is, therefore, expected to be accurate in the vertical direction and not as good in the horizontal direction. This is borne out in practice.

2. XORstep: Next the XOR image is computed. This is done by XOR'ing corresponding pixels. A match error EXOR may be computed by finding the number of white pixels in the XOR image. The XOR image match error is in general not accurate enough for matching. Notice that XOR images may consist of either isolated pixels or pixels in a blob. The error measure computed above gives equal weight to both. However, an isolated pixel in the XOR image may be due to noise while a blob may be due to a major mismatch. Therefore, blobs should be given more weight. This can be done by using an Euclidean distance mapping.

3. EDMstep: An Euclidean distance mapping [DAN80] is computed from the XOR image by assigning to each white pixel in the image, its minimum distance to a black pixel. Thus a white pixel inside a blob will get a larger distance than an isolated white pixel. An error measure EEDM can now be computed by adding up the distance measures for each pixel.

4. Although the approximate translation has been computed using step 1, this may not be accurate and may need to be fine-tuned. Thus steps 2 and 3 are repeated while sampling the translation space in both x and y. A minimum error measure EEDMmin is computed over all the translation samples.

IV.EXPERIMENTAL EVALUATION

A.Dataset Description

For testing the proposed keyword spotting method uses the George Washington database (GW DB)

*GW DB***:** The GW Data set consists of 20 pages of letters, orders, and instructions of George Washington from 1755. The pages originate from a large collection with a variety of images, the quality of which ranges from clean to very difficult to read. The selected pages we use are relatively clean. The text is part of a larger corpus, written not only by George Washington but also by some of his associates. It inhibits some variations in writing style. However, the writing on the pages we consider is fairly similar. The considered pages include 4,894 words on 675 text lines. The GWDB contains the same pages as the one, but we found the automatically segmented and extracted words to be too erroneous. Focusing on keyword spotting rather than document image preprocessing in this paper, we manually segmented the data set into individual words. Hence, there is a slight difference in the number of words and word classes.Following Fig.2 shows this.

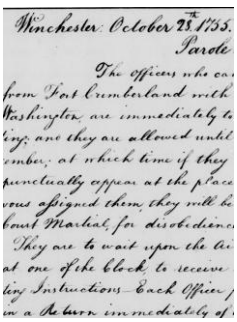

Fig.2(a) Input image

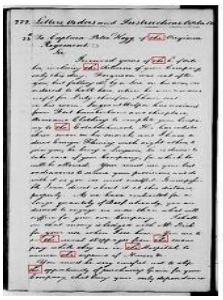

Fig 2(b) Word Spotted

B.Performance Matrices

The proposed method of word spotting is compared with word spotting using HMM model.The technique proposed here can outperform than HMM word-based recognition technique in terms of word error rates. The results reported here focus on word error rates (WER).Three or Four letter words correctly spotted without any error. Lavrenko et al. [16] trained an HMM model on 19 pages of the dataset used in this paper and tested on the remaining page. With 20-fold cross validation they obtained a word error rate of 41% (excluding out of vocabulary terms) and 50% (when including out of vocabulary terms). By including bigrams from a Jefferson corpus and a Washington corpus (excluding the test set) they

reduced the WER to 35% and 45% respectively. We note that the best approach in this paper has an even lower WER, showing that the word spotting approach is quite competitive.

V. CONCLUSION AND FUTURE WORK

Wordspotting appears as an alternative to the seemingly obvious recognize-then-retrieve approach to historical manuscript retrieval.The present work proposes one type of Information retrieval from handwritten word images.Here input query is in text form. With this method keywords are spotted in the all images for a given query keyword.Comparative analysis of the proposed system provides better performance result rather than existing system. The results obtained in this work, although preliminary are encouraging to continue with a further research in different directions. Our further research has to be oriented to big databases of handwritten images.

REFERENCES

[1] T.M. Rath and R. Manmatha, "Word Spotting for Historical Documents," Int'l J. Document Analysis and Recognition, vol. 9, pp. 139-152, 2007.

[2] T. Rath and R. Manmatha, "Features for word spotting in historical manuscripts," ICDAR, pp. 218–222, 2003.

[3] H. Cao and V. Govindaraju, "Template-free word spotting in low-quality manuscripts," ICDAR, pp. 392–396, February 2007.

[4] T. Adamek, N. E. Connor, and A. F. Smeaton, "Word matching using single closed contours for indexing handwritten historical documents," IJDAR, vol. 9, no. 2, pp. 153–165, 2007.

[5] M.Rusi˜nol, D.Aldavert, R.Toledo, and J.Llad´os, "Browsing heterogeneous document collections by a segmentation-free word spotting method," ICDAR, pp. 63–67, 2011.

[6] J. A. Rodr´ıguez-Serrano, F. Perronnin, and J. Llados, "A similarity measure between vector sequences with application to handwritten word image retrieval," CVPR, pp. 1722–1729, August 2009.

[7] C. Choisy, "Dynamic handwritten keyword spotting based on the nshphmm," Proceedings of the Ninth ICDAR, vol. 1, pp. 242–246, 2007.

[8] J. A. Rodr´ıguez-Serrano and F. Perronnin, "Handwritten word-spotting using hidden markov models and universal vocabularies," Pattern Recognition, pp. 2106–2116, February 2009.

[9] V. Frinken, A. Fischer, R. Manmatha, and H. Bunke, "A novel word spotting method based on recurrent neural networks," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 34, no. 2, pp. 211–224, 2012.

[10] S.Thomas, C.Chatelain, L.Heutte, and T.Paquet, "An information extraction model for unconstrained handwritten documents," ICPR, Istanbul, Turkey, pp. 1–4, 2010.

[11] A. Fischer, A. Keller, V. Frinken, and H. Bunke, "Lexicon-free handwritten word spotting using character hmms," Pattern Recognition Letters, vol. 33, no. 7, pp. 934 – 942, 2012.

[12] S. Weshah, G. Kumar, and V. Govindaraju, "Script independent word spotting in offline handwritten documents based on hidden markov models," In proceeding of: The 13th International Conference on Frontiers in Handwriting Recognition, (ICFHR 2012), 2012.

[13] R.Manmatha and W.B. Croft, Word Spotting: Indexing Handwritten Archives, ch. 3, pp. 43-64. MIT Press, 1997

[14] U.-V. Marti and H. Bunke, "Using a Statistical Language Model to Improve the Performance of an HMM-Based Cursive Handwriting Recognition System," Int'l J. Pattern Recognition and Artificial Intelligence, vol. 15, pp. 65-90, 2001.

[15] F. Perronnin and J. Rodriguez-Serrano, "Fisher Kernels for Handwritten Word-spotting," in *10th Int'l Conf. on Document Analysis and Recognition*, vol. 1, 2009, pp. 106–110.

[16] Lavrenko, V., Rath, T. M., and Manmatha, R. Holistic word recognition for handwritten historical documents. In Proc. of the Int'l Workshop on Document Image Analysis for Libraries (Palo Alto, CA, January 23-24 2004), .278-287.