ISSN (Online) : 2319 - 8753
ISSN (Print) : 2347 - 6710

International Journal of Innovative Research in Science, Engineering and Technology

An ISO 3297: 2007 Certified Organization Volume 3, Special Issue 5, July 2014

International Conference On Innovations & Advances In Science, Engineering And Technology [IC - IASET 2014]

Organized by

Toc H Institute of Science & Technology, Arakunnam, Kerala, India during 16th -18th July -2014

An Advanced Approach for Text Query Searching and Word Spotting In Word Images

Haritha V R[1], Sreeram S[2]

PG Student [CSE], Dept. of CSE, MEA Engineering College, Perinthalmanna, Kerala, India [1]

HOD, Dept. of CSE, MEA Engineering College, Perinthalmanna, Kerala, India [2]

ABSTRACT: Word-spotting refers to the problem of detecting specific keywords in document images. Here we focus on handwritten word images. Keyword spotting in handwritten image document in the existing work is based upon BLSTM Neural Networks which consist of two parts. First part is preprocessing phase, performed by the neural network. It maps each position of an input sequence to a vector, indicating the probability of each character possibly being written at that position. The second part, called the CTC Token Passing algorithm, takes this sequence of letter probabilities, as well as a dictionary and a language model, as its input and computes a likely sequence of words. By extending this work, the present work proposes Information retrieval and information (text) extraction methods from all handwritten documents of images. In Information retrieval approach the input query is text format .The text is matched with template character then the query image is created from template characters. This proposed approach provides an efficient way of searching text like queries in document images. The text extraction from the images includes thresholding, segmentation, edge detection and text extraction algorithm. The experimental results show the performance of the proposed algorithms achieves higher accuracy rates than existing approaches.

KEYWORDS: Keyword spotting, Offline handwriting, Historical documents, Neural network, BLSTM.

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 digitized images of documents. There are lots of historical handwritten documents with information that can be used for several studies and projects. 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. 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) that are close to the query in terms of a specific dissimilarity measure. 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.

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.

Handwriting recognition is far from easy. A common complaint and excuse of people is that they couldn't read their own handwriting. That makes us ask ourselves the question: If people sometimes can't read their own handwriting, with which

ISSN (Online) : 2319 - 8753
ISSN (Print) : 2347 - 6710

International Journal of Innovative Research in Science, Engineering and Technology

An ISO 3297: 2007 Certified Organization Volume 3, Special Issue 5, July 2014

International Conference On Innovations & Advances In Science, Engineering And Technology [IC - IASET 2014]

Organized by

Toc H Institute of Science & Technology, Arakunnam, Kerala, India during 16th -18th July -2014

they are quite familiar, what chance does a computer have? Fortunately, there are powerful tools that can be used that are easily implementable on a computer. A very useful one for handwriting recognition, and one that is used in several recognizers, is a neural network.

Neural networks are richly connected networks of simple computational elements. The fundamental tenet of computation with neural networks is that such networks can carry out complex cognitive and computational tasks. In addition, one of the tasks at which neural networks excel is the classification of input data into one of several groups or categories. This ability is one of the main reasons neural networks are used for this purpose. In addition, neural networks fare well with noisy input, making them even more attractive.

II. RELATED WORK

Current approaches to word spotting can be split into two categories, viz. query-by-example (QBE) and query-by string (QBS). With the former approach, all instances of the search word in the training set are compared with all word images in the test set. Among the most popular approaches in this category are dynamic time warping (DTW) [3], [4], [5] and classification using global features [6], [7]. Algorithms based on QBE suffer from the drawback that they can only find words appearing in the training set. The latter approach of QBS models the key words according to single characters in the training set and searches for sequences of these characters in the test set [8], [9].

Recently, keyword spotting systems that are modified versions of handwriting recognition systems have received increasing attention. In [9], [10], [11], hidden Markov models are used to find the words to be searched. In [12], a novel approach using bidirectional long short-term (BLSTM) neural networks is proposed. However, the performance of the neural network based keyword spotting system depends crucially on the amount of training data.

III. PROPOSED SYSTEM

The proposed system solves all the problems, and also provides more flexible keyword spotting like text based queries which is case insensitive while the image based queries are case sensitive. The proposed work uses Information retrieval method and Text extraction method of all images for keyword spotting. This work can be done by extending the keyword spotting method for handwritten text based on BLSTM Neural Networks [1] and CTC Token Passing algorithm. The proposed system spots keywords in images based on given query keyword. Unlike existing work, the query keyword can be given in the form of text of characters. Based on query keyword, the keyword spotting can be done by using BLSTM Neural Networks and CTC Token Passing algorithm [1] for all images. But the line text is extracted from the matched keyword. The text extraction steps proposed in this method as follows

The methods of Information (text) extraction from the images can be done by using thresholding, segmentation, edge detection and comparison algorithm

The Main Steps of BLSTM Word spotting System

1. A scanned grey level image of the document is obtained
2. The image is first reduced by half by Gaussian filtering and sub sampling.
3. The reduced image is then binarized by thresholding the image (fig). [14]
4. The binary image is now segmented into text lines. This is done by a process of smoothing and thresholding ([13]).

ISSN (Online) : 2319 - 8753
ISSN (Print) : 2347 - 6710

International Journal of Innovative Research in Science, Engineering and Technology

An ISO 3297: 2007 Certified Organization Volume 3, Special Issue 5, July 2014

International Conference On Innovations & Advances In Science, Engineering And Technology [IC - IASET 2014]

Organized by

Toc H Institute of Science & Technology, Arakunnam, Kerala, India during 16th -18th July -2014

5. From each line, a sequence of feature vectors is extracted [15]. It maps each position of an input sequence to a vector, indicating the probability of each character possibly being written at that position.
6. Then CTC Token Passing algorithm takes this sequence of letter probabilities, as well as a dictionary and a language model, as its input and computes a likely sequence of words.[1]

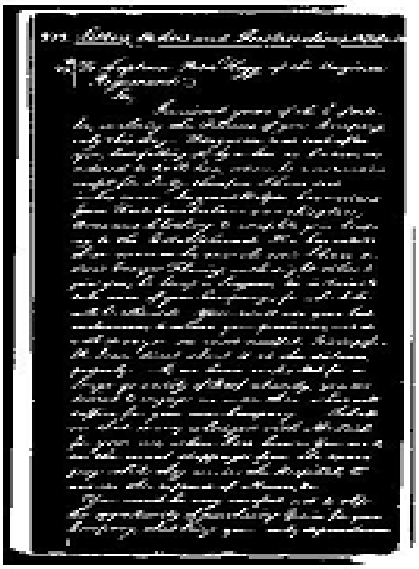

Fig: Before Segmentation of a page

Fig: word'the'spotted in GW image

3.4 Information retrieval from all document images

In this module, text query is given as input. Based on the text query, the query template character has been generated. At first, from the all document, template character has been taken and stored in the database. When the text query is given, the system matches the text query with the saved template characters. Once the text query matches with the template characters, the template can be retrieved from the database. From this, information of queried text document is retrieved and analysed for all document images.

A step for information retrieval is defined as follows:

Step 1: Input the text query
Step 2: Extract the templates from the document and store in a database
Step 3: Perform matching of text query with the templates stored in database
Step 4: If the text matches with the Template character

ISSN (Online) : 2319 - 8753
ISSN (Print) : 2347 - 6710

International Journal of Innovative Research in Science, Engineering and Technology

An ISO 3297: 2007 Certified Organization Volume 3, Special Issue 5, July 2014

International Conference On Innovations & Advances In Science, Engineering And Technology [IC - IASET 2014]

Organized by

Toc H Institute of Science & Technology, Arakunnam, Kerala, India during 16th -18th July -2014

Retrieve the template character for the text

Else

Go to step 3

Step 5: Return the retrieved template character

Step 6: Exit

3.5 Information (text) Extraction

In this module the text information from the images are extracted by using the following algorithms.

3.5.1Segmentation Algorithm

1. Get the 2-dimensional arrays containing pixel intensity of the pixels in the image.
2. Scan the array row wise and identify a pair of boundaries – the upper and the lower line boundaries for each line of text in image and store in a separate array.
3. Within each pair of line boundaries identify the word boundaries whenever a few columns with all **1**'s in them is encountered and store in another 2-dimensional array.
4. Within each of these identified word boundaries determine and store the character boundaries in a 2-dimensional array, whenever one or more columns of the pixel intensity values having all **1**'s is encountered.

3.5.2 Edge-detection Algorithm

1. For each character, obtain the segmentation boundaries and the pixel mapping from the 2-dimensional array.
2. Identify the topmost left corner text pixel from this pixel mapping to start the edge detection. Set the current pixel position as the start pixel position.
3. Store the direction of movement as right.
4. Trace the whole boundary of the character using **8-Neighbourhood technique** in a clockwise sequence by repeating the following steps:

   4.1) Store the current pixel position as (x, y) co-ordinate entry in two 2-dimensional arrays- one for the for the input and the other for template respectively.

   4.2) Identify all possible neighborhood pixels in the text for edge tracing.

   4.3) Move to the next possible adjacent pixel in clockwise sequence from the current position according to the direction of movement.

   4.4) Store the last direction of movement.

   4.5) Break if the current pixel position reaches start pixel.

5. Repeat steps from (1) to (4) with intermediate delimiters until all characters are processed.

ISSN (Online) : 2319 - 8753
ISSN (Print) : 2347 - 6710

International Journal of Innovative Research in Science, Engineering and Technology

An ISO 3297: 2007 Certified Organization Volume 3, Special Issue 5, July 2014

International Conference On Innovations & Advances In Science, Engineering And Technology [IC - IASET 2014]

Organized by

Toc H Institute of Science & Technology, Arakunnam, Kerala, India during 16th -18th July -2014

3.5.3 Comparison Algorithm:

When the edge-detected arrays for the templates and input image are ready, do the following:

1. Store the (x, y) co-ordinate values between the delimiters in a separate 2-dimensional array to represent a single character from the input image.
2. Use a separate variable for count of mismatches and set it to some high value.
3. Until the character boundaries in the template array is exhausted:

   3.1.a) Compare each of the [x, y) co-ordinate positions stored in the template array with that of input array.
   b) Allowance of +1 or -1 pixel is allowed in each pixel comparison
   c) If it does not match increment the mismatches by one.
   3.2 Update the mismatches value if the current mismatch value for the template character is lesser than the existing value.

4. Choose the character from the template corresponding to the least mismatches, as the selected character and write it to the text file.
5. Add delimiters for indicating next character, word and line.

Repeat the steps from (1) to (5) until there are no more boundary values available in the array used for the input text

3.5.4 Text extraction algorithm

1. Create the template images according to the given specifications.
2. Select the image from which the text has to be extracted.
3. Get the colour of the printed text embedded in the image (i.e., the text pixel intensity).
4. Disintegrate the whole image into pixels and store their intensity values in an array.
5. Set the threshold value for choosing the text pixels from the image according to the colour chosen (i.e., set the colour of the text pixels as **8** and all other pixel values as **1**).
6. Once the threshold value has been set, the pixel intensity values are stored in a 2-dimensional array.
7. Apply segmentation algorithm.
8. Apply edge detection algorithm.
9. Apply comparison algorithm.
10. Write the extracted text to a text file and save the file.

ISSN (Online) : 2319 - 8753
ISSN (Print) : 2347 - 6710

International Journal of Innovative Research in Science, Engineering and Technology

An ISO 3297: 2007 Certified Organization Volume 3, Special Issue 5, July 2014

International Conference On Innovations & Advances In Science, Engineering And Technology [IC - IASET 2014]

Organized by

Toc H Institute of Science & Technology, Arakunnam, Kerala, India during 16th -18th July -2014

IV. EXPERIMENTAL EVALUATION

4.1 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.

4.2 Performance Matrices

The proposed method of BLSTM-NN with Information retrieval and extraction approach is compared with the BLSTM-NN. Comparison with BLSTM-NN in the application area of keyword spotting with Precision versus Recall diagrams on ground truth databases reveal that the proposed approach achieves better scores.

Precision

The precision rate is defined as the ratio of the number of relevant images retrieved and total number of images in the collection.

Precision=|{relevant images}∩{retrieved images}|/|{retrieved images}|

Recall

Recall rate is defined as the ratio of number of relevant images retrieved and to the total number of relevant images in the collection.

Recall=|{relevant images}∩{retrieved images}|/{relevant images}

The experimental results for proposed system are plotted on graphs based on these formulas. The graphs shown in the figures below give better analysis perspective on the Information retrieval and extraction task.

ISSN (Online) : 2319 - 8753
ISSN (Print) : 2347 - 6710

International Journal of Innovative Research in Science, Engineering and Technology

An ISO 3297: 2007 Certified Organization Volume 3, Special Issue 5, July 2014

International Conference On Innovations & Advances In Science, Engineering And Technology [IC - IASET 2014]

Organized by

Toc H Institute of Science & Technology, Arakunnam, Kerala, India during 16th -18th July -2014

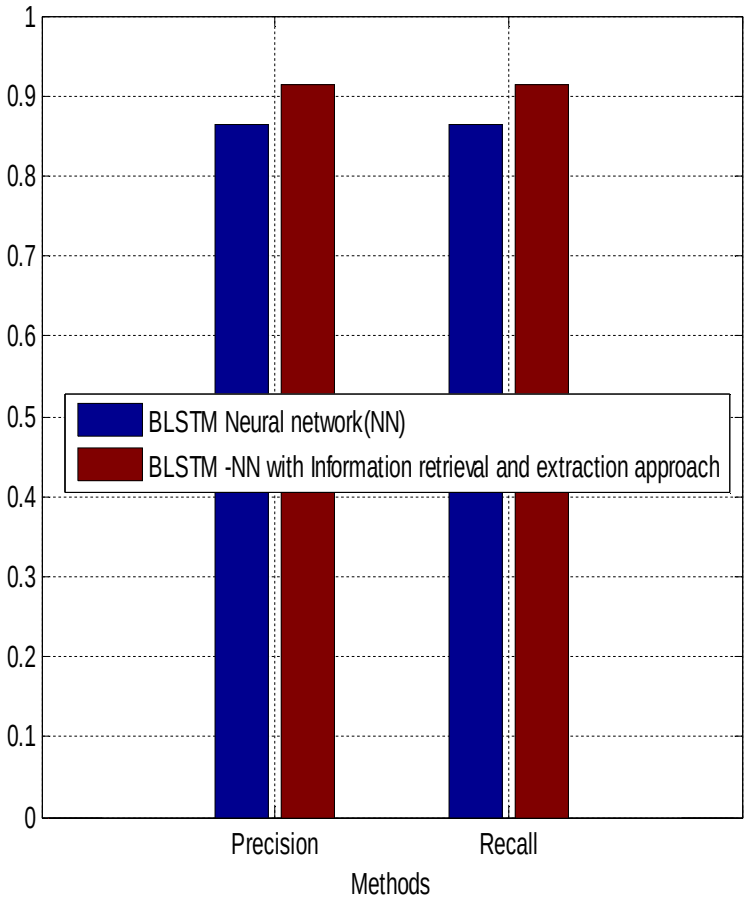

Fig: 5.1 Precision-Recall Comparison graphs

The above graph in figure 5.1 compare the Precision-Recall parameter between BLSTM-NN and BLSTM-NN with Information retrieval and extraction approach. These measures are mathematically calculated by using formula. This graph shows the Precision-Recall rate of BLSTM-NN and BLSTM-NN with Information retrieval and extraction approach. In this graph X-axis will be methods such as Precision and Recall for BLSTM-NN and BLSTM-NN with Information retrieval and extraction approach and Y-axis will be value for the precision and recall. From the graph can see that, Precision-Recall of the system is reduced somewhat in BLSTM-NN than the BLSTM-NN with Information retrieval and extraction approach. From this graph we can say that the Precision-Recall rate of BLSTM-NN with Information retrieval and extraction approach is increased which will be the best one.

ISSN (Online) : 2319 - 8753
ISSN (Print) : 2347 - 6710

International Journal of Innovative Research in Science, Engineering and Technology

An ISO 3297: 2007 Certified Organization Volume 3, Special Issue 5, July 2014

International Conference On Innovations & Advances In Science, Engineering And Technology [IC - IASET 2014]

Organized by

Toc H Institute of Science & Technology, Arakunnam, Kerala, India during 16th -18th July -2014

V.CONCLUSION

The present work proposes Information retrieval and text information extraction .of the keyword spotted images. Keyword spotting approach using bidirectional long short-term neural networks (BLSTM NN) in combination with a modified version of the Connectionist Token Passing algorithm. With this method keywords are spotted in the all images for a given query keyword. This method efficiently ranks the images with the highest matched keywords based on query keyword. The proposed system efficiently extracted the information (text) of all images by using segmentation, edge detection and text extraction algorithm. The results from this algorithm are refined by using Thresholding mechanism for efficient information Extraction. Comparative analysis of the proposed system provides better performance result rather than existing system.

REFERENCES

[1] V Frinken, **A** Fischer, R Manmatha, Horst Bunke,"A Novel Word Spotting Method Based on Recurrent neural networks" IEEE Trans . Pattern Analysis and Machine Intelligence,, VOL. 34, NO. 2, Feb 2012
[2] A. Graves, M. Liwicki, S. Ferna´ndez, R. Bertolami, H. Bunke, and J. Schmidhuber, "A Novel Connectionist System for Unconstrained Handwriting Recognition," IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 31, no. 5, pp. 855-868, May 2009
[3] A. Kołcz, J. Alspector, M. F. Augusteijn, R. Carlson, and G. V. Popescu, "A Line-Oriented Approach to Word Spotting in Handwritten Documents," *Pattern Analysis and Applications*,vol. 3, pp. 153–168, 2000.
[4] R. Manmatha and T. M. Rath, "Indexing of Handwritten Historical Documents - Recent Progress," in *Symposium on Document Image Understanding Technology*, 2003, pp. 77–85.
[5] T. M. Rath and R. Manmatha, "Word Image Matching Using Dynamic Time Warping," in *Computer Vision and Pattern Recognition*, vol. 2, 2003, pp. 521–527.
[6] E. Ataer and P. Duygulu, "Matching Ottoman Words: An Image Retrieval Approach to Historical Document Indexing," in *6th Int'l Conf. on Image and Video Retrieval*, 2007, pp.341–347.
[7] Y. Leydier, F. Lebourgeois, and H. Emptoz, "Text Search for Medieval Manuscript Images," *Pattern Recognition*, vol. 40, pp. 3552–3567, 2007.
[8] H. Cao and V. Govindaraju, "Template-free Word Spotting in Low-Quality Manuscripts," in *6th Int'l Conf. on Advances in Pattern Recognition*, 2007.
[9] J. Edwards, Y. Whye, T. David, F. Roger, B. M. Maire, and G. Vesom, "Making Latin Manuscripts Searchable using gHMM's," in *Advances in Neural Information Processing Systems (NIPS) 17*. MIT Press, 2004, pp. 385–392.
[10] H. Jiang and X. Li, "Incorporating training errors for large margin hmms under semi-definite programming framework," *Int'l Conf. on Acoustics, Speech and Signal Processing*, vol. 4,pp. 629–632, April 2007.
[11] 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.
[12] V. Frinken, A. Fischer, and H. Bunke, "A Novel Word Spotting Algorithm Using Bidirectional Long Short-Term Memory Neural Networks," in *4th Workshop on ArtificialNeural Networks in Pattern Recognition*, 2010.
[13] R. Manmatha, Chengfeng Han, E. M. Riseman, and W. B. Croft. Indexing handwriting using word matching. In *Dig- ital Libraries .96: 1st ACM International Conference on Digital Libraries*, 1996.
[14] R. Manmatha and W.B. Croft, Word Spotting: Indexing Handwritten Archives, ch. 3, pp. 43-64. MIT Press, 1997
[15] 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 andArtificial Intelligence, vol. 15, pp. 65-90, 2001.