A NOVEL APPROACH FOR WORD RETRIEVAL FROM DEVANAGARI DOCUMENT IMAGES

Blessy Varghese and Sharvari Govilkar

Department of Computer Engineering, University of Mumbai, PIIT, New Panvel, India

ABSTRACT

Large amount of information is lying dormant in historical documents and manuscripts. This information would go futile if not stored in digital form. Searching some relevant information from these scanned images would ideally require converting these document images to text form by doing optical character recognition (OCR). For indigenous scripts of India, there are very few OCRs that can successfully recognize printed text images of varying quality, size, style and font. An alternate approach using word spotting can be effective to access large collections of document images. We propose a word spotting technique based on codes for matching the word images of Devanagari script. The shape information is utilised for generating integer codes for words in the document image and these codes are matched for final retrieval of relevant documents. The technique is illustrated using Marathi document images.

KEYWORDS

Word spotting, document image retrieval, keyword spotting, searching and retrieval, word image, Marathi document, word matching, Devanagari script.

1. INTRODUCTION

Searching for information from document images containing predominantly text, the traditional Information Retrieval (IR) approach using keywords is often used. For these document images, conventional document image processing techniques can be easily utilized for the purpose. For instance, many document image retrieval systems first convert the document images into their text format using OCR techniques and then apply text IR strategies over the converted text documents. Several commercial systems have been developed for this purpose.

All these systems convert the document images into their electronic representations to facilitate text retrieval. However, high costs and poor quality of document images often prohibit complete conversion using OCR. In such conditions, it is better to explore techniques using image features in order to retrieve document images containing text. Generally the recognition accuracy requirements for document image retrieval are considerably lower than that for document image processing tasks. A document image processing system will analyse different text regions, understand the relationships among them, and then converting them into machine-readable text using OCR.

On the other hand, a document image retrieval system [2] asks whether a document image contains particular words which are of interest to the user, ignoring other unrelated words. Thus, a document image retrieval system answers “yes” or “no” to the user’s query, instead of exact recognition of characters or words. This is sometimes known as “keyword spotting” or simply “word spotting” with no need for correct and complete character recognition but by directly characterizing image document features at the character, word or line level.

DOI: 10.5121/ijnlc.2015.4402

The paper presents a novel approach for word retrieval from Marathi document image. Related work done using several techniques and past literature is discussed in section 2. The proposed approach is discussed in detail in section 3. Finally, the last section 4 concludes the paper.

2. LITERATURE SURVEY

The In this section, we cite the relevant past literature that utilizes the various techniques for content based document image retrieval.

Toni M. Rath, et.al. [1] presents an algorithm for matching handwritten words in noisy historical documents. The segmented word images are pre-processed to create sets of 1-dimensional features, which are then compared using dynamic time warping (DTW). Features like projection profile, word-profiles and background-ink transitions are extracted from the word images and considered for matching. Experimental results on two different data sets from George Washington collection are given.

Sargur N. Srihari, et.al. [3] describes a system for spotting words in scanned document images in three scripts, Devanagari, Arabic and Latin. The user gives a query which can be either a word image or text. The candidate words that are searched in the documents are retrieved and ranked, where the ranking criterion is a similarity score (correlation) based on global word shape features (Geometrical, Structural, Concavity features). When the query is given in text form, the text query is mapped into a number of query word images which can later be matched with the indexed word images. Handwritten samples of a word are obtained from an indexed (segmented) set of documents. These indexed documents contain the truth (English equivalent) for every word image.

Anurag Bhardwaj, et.al. [4] discusses a method for script independent word spotting in multilingual handwritten and machine printed documents. The Indexer performs indexing of all word images present in the document image corpus by extracting Moment Based features from word images and storing them as index. The Similarity Matcher returns a ranked list of word images which are most similar to the query based on a cosine similarity metric. The performance of the system is seen to be superior on printed text than on handwritten text. Experiments are reported on documents of three different languages: English, Hindi and Sanskrit.

Million Meshesha, et.al. [5] The author designed a novel partial matching algorithm for morphological matching of word form variants in a language by modifying the DTW algorithm. Local features like projection profiles, transition profiles, etc. are extracted by scanning vertical strips of the word image. Performance analysis of the proposed approach on English, Amharic and Hindi documents is presented. The introduction of DTW-based partial matching scheme enables to control morphological variants of a word.

Anurag Bhardwaj, et.al. [7] discusses two methods for keyword spotting in printed Sanskrit documents, one which is recognition based and other which is recognition free. The first approach is script specific which uses a Devanagari OCR based Block Adjacency Graph (BAG) scheme for word recognition. It includes a BAG based technique that uses a graph to maintain the overall character structure. The second approach is a moment based word matching technique which maintains a script invariant representation of all word images. Word matching is performed using the cosine similarity. A relevance feedback technique is employed to refine the word spotting results.

Shuyong Bai, et.al [8] proposed an algorithm based on Word Shape Coding and matching algorithm. The page is segmented into words and for each word, features are detected and converted into word shape codes (WSC). The method is experimented on English documents and so different codes are assigned to the 52 characters of the language. These character codes are concatenated to form the word codes which are used for matching with the query word. A recurrence equation is used as the matching measure. The approach is demonstrated on printed English documents belonging to two different datasets.

Arundhati Tarafdar, et.al [9] proposes a shape code based word-image matching technique for retrieval of multilingual documents in Indian languages. Word images are represented using primitive shape codes using various word shape information like loop positions, crossing points, extreme points, etc. Candidate words are chosen from document images based on features like aspect ratio, etc. and represented using the shape codes. An inexact string matching technique is applied for matching. Printed documents of Bangla, Devanagari and Gurumukhi scripts were considered for the experiments.

Safwan Wshah, et.al. [10] presents line-based keyword spotting based on Hidden Markov Model (HMM) which simulates the keywords in model space as sequence of character models and uses filler models for background or non-keyword text. The use of filler models improves the retrieval result as non-relevant words are handled appropriately by reducing their cost from the overall cost. The method is demonstrated on handwritten documents of English, Arabic and Devanagari.

Volkmar Frinken, et.al. [11] presents a keyword spotting method based on BLSTM neural network and CTC token passing algorithm. The pre-processing phase performed by the neural network maps each position of an input sequence to a vector, indicating the probability of each character possibly being written at that position. The CTC algorithm generates a token for every character and every position in the text line which stores the probability of that character being present at that position together with the probability of the best path from the beginning to that position. It was tested on three datasets of English handwritten documents and gave promising results.

Traditional approach for word spotting includes extraction of feature vectors based on profile features, GSC features, Moment based features, etc. from the word images followed by matching algorithms like Euclidean distance, Cosine similarity, Correlation measure, DTW algorithm, etc. Hidden Markov Models (HMMs) are also used for word spotting. As per observations in [11] novel approach based on Neural Network are more powerful techniques as compared to the traditional word spotting approaches but is more suited to complex problems like recognition of handwritten documents.

3. PROPOSED APPROACH

The idea is to search the document image for the input query word specified by the user by extracting image-based features from the word image and matching it with the document image. The word images are represented using integer codes devised from the shape information of the word, and candidate words are matched with the query word.

We propose a system where the system accepts a textual query in Marathi language from the user along with a Marathi printed document image in which the keyword has to be searched. The textual query is converted to image and the document image is also segmented into word-images. Word-shape codes are generated from these word-images, and search is carried out for retrieval of relevant documents by matching the word-shape codes of the query word-image with the word-images of the document.

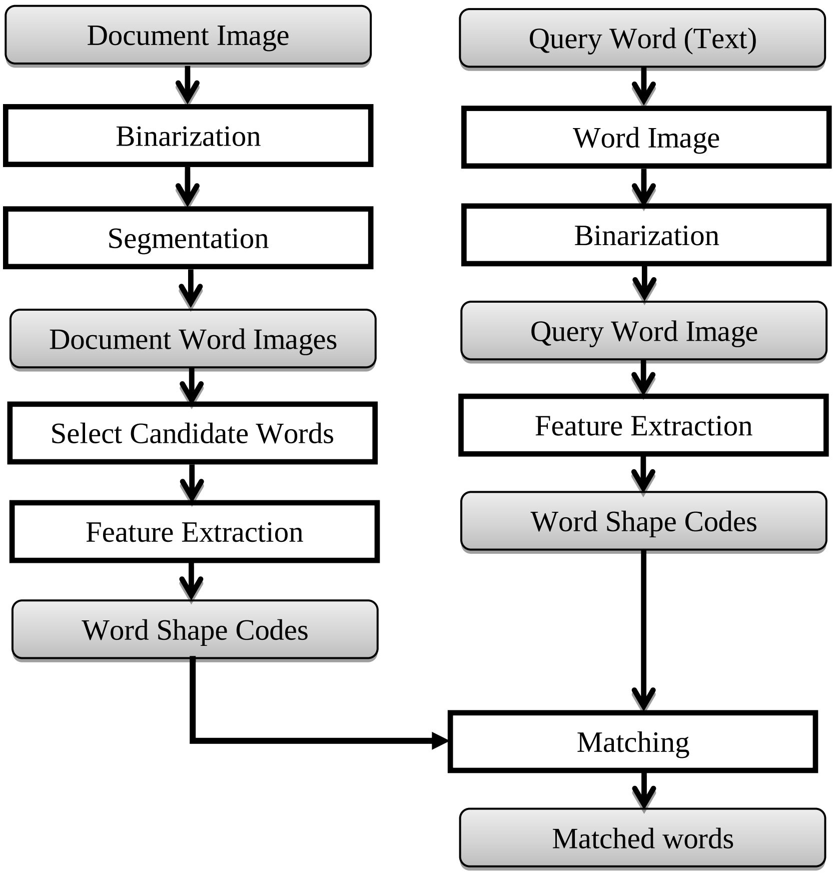

Figure 1: Proposed Approach.

The classical approaches utilize the (profile) features as a whole to construct the feature vector which results in an extended and more complex encoding technique. We rather develop integer codes based on the shape and size of the word which gives rise to a method which is simple as well as requires comparatively minimal space.

The proposed approach consists of following phases:
1. Pre-processing phase
2. Feature Extraction and Encoding phase
3. Matching phase

The first step is to pre-process the input image by applying binarization, segmentation, etc. From each word image, integer codes are generated by extracting the relevant features. These word-shape integer codes or simply integer codes are then used for matching.

3.1. Pre-processing Phase

This phase includes the pre-processing activities required to be performed for the input images. The input image has to be binarized, segmented into word images, and then features are extracted from these word images.

Query word taken as input from the user in text form is converted into word image and then the activities for pre-processing of the word image takes place. Both the document image as well as the word image is first binarized, document image is then segmented to give word images. Features are extracted from these word images and encoded to form the word shape integer codes which are then used for matching.

3.2. Feature Extraction and Encoding phase

Marathi text is written in three zones: upper zone, middle zone and lower zone. The upper and lower zones comprise the modifiers and most of the information is present in the middle zone. For each feature, an integer code is generated which is then used for matching.

Following different codes are generated for the word images:

3.2.1. Extreme point's code

The header line is removed to get individual independent elements which can be alphabets, joint letters, etc. After the header line is removed, middle zone of each word is divided into two parts. Now, from each disconnected element of the upper and lower part, the extreme points are found. For the element lying on the upper middle zone, compute upper extreme points using profile information from the top and similarly for the lower part from the bottom.

According to the location of the extreme points, six codes can be computed. Four codes if the point lies in one of the zones (upper, upper middle, lower middle, lower) and two codes if the extreme point lies on the header line or base line. Codes are assigned numbering from 1 to 6 for each point, from top to bottom [9]. These are:

- Code 1 - point lies in upper zone
- Code 2 - point lies on header line
- Code 3 - point lies in upper middle zone
- Code 4 - point lies in lower middle zone
- Code 5 - point lies on base line
- Code 6 - point lies in lower zone

3.2.2. Vertical bar code

Marathi text has alphabets which mostly contains a vertical bar like structure. The position of these vertical bars in the word is used to generate a new code. The column value i.e. the x-coordinate is normalized using the word width to give a single integer for each vertical bar and these integers are sequentially put together from left to right to generate the final code.

3.2.3. Crossing based code

As Marathi words mostly contain vertical bar like structures, the word is divided into smaller regions using these vertical bars. If there are ‘n’ vertical bars, we get ‘n+1’ regions (region from start of word to first vertical bar, regions between the vertical bars and region from last vertical bar to the end of the word). Each of these regions is further divided into four parts by three equidistant lines and if the line is crossed by any element is recorded. The number of crossings is recorded for each region (maximum will be 4 and minimum will be 0) and sequentially put together starting from left to give the final code.

3.2.4. Loop based code

The number of closed regions or loops, their location in each word and the size of each loop is different. Using this feature, two different codes can be computed. For each closed region, the centre of gravity (CG) is found and its x-coordinate is normalized using the width of the word. This is the first code using loop. The sequence of loop sizes from left to right will be different from one word to another even though the size of loops of particular character in a font is fixed. This information is useful for encoding. The height and breadth of all loops found in a word are found and the heights of all such loops are normalized with respect to maximum height found among them and similarly for breadth also.

3.2.5. Background element coding

Background can provide much useful information that helps in word spotting scheme. Each background element can be considered approximately as a character in the word which is accomplished without explicit character segmentation. The background between two elements is detected and run length information is used. Closed regions are not considered for this encoding. Horizontal scanning for each background element is done to find the maximum (max_i) and minimum (min_i) runs among all the runs in the upper part of middle zone [9] and similarly for lower middle zone. The code is normalized using the maximum width (mmax) found among all max_i values found in a word image. The code is generated based on this maximum and minimum run information of the background elements. So we get two code values for each element.

Features are extracted from the word images in vertical stripes by sliding window across the word image and encoded into integer codes. These integer shape codes are then considered for matching to get the final result.

3.3. Matching phase

For fast retrieval, first some candidate words are found and matching is done on these words. Candidate words are selected from the document images based on properties like aspect ratio, number of loops, number of background components, etc. A threshold is set for each of the properties to ensure that no relevant words are omitted. The selection of candidate words reduces the overall processing time as the number of word images to be processed is decreased.

The word shape codes are generated for the candidate words and for the query word image which is used for matching. DTW (Dynamic Time Warping) [6] algorithm is used for matching the word shape codes. This algorithm is slightly modified to work for partial matching [5]. It is not mandatory that all the codes should match for retrieving the given query word. The minimum number of codes that should match so as to consider the word in the result set can be decided by testing it with different samples and according to the extent of correctness expected for the respective application.

4. CONCLUSIONS

Word spotting is an effective technique which can be used for retrieving relevant information from document images where OCR doesn't give promising results. We propose a system where the features are extracted from the shape information of the word images of Devanagari script and integer codes are generated for each word image which is then used for matching using the modified DTW algorithm. The matched set of words is then sorted according to the similarity score to display the final set of documents.

REFERENCES

[1] Rath, Toni M., and Raghavan Manmatha. "Word image matching using dynamic time warping." Computer Vision and Pattern Recognition, 2003, Proceedings.2003 IEEE Computer Society Conference on Vol.2 IEEE, 2003.

[2] Lu, Yue, Li Zhang, and Chew Lim Tan. "Retrieving imaged documents in digital libraries based on word image coding." Document Image Analysis for Libraries, 2004. Proceedings. First International Workshop on. IEEE, 2004.

[3] Srihari, Sargur N., et al. "Spotting words in Latin, Devanagari and Arabic scripts." VIVEK-BOMBAY- 16.3 (2006): 2.

[4] Bhardwaj, Anurag, Damien Jose, and Venu Govindaraju. "Script Independent Word Spotting in Multilingual Documents". IJCNLP. 2008.

[5] Meshesha, Million, and C. V. Jawahar. "Matching word images for content-based retrieval from printed document images. "International Journal of Document Analysis and Recognition (IJDAR) 11.1 (2008): 29-38.

[6] Senin, Pavel. "Dynamic time warping algorithm review." Information and Computer Science Department University of Hawaii at Manoa Honolulu, USA (2008): 1-23.

[7] Bhardwaj, Anurag, SrirangarajSetlur, and Venu Govindaraju. "Keyword spotting techniques for Sanskrit documents." Sanskrit Computational Linguistics. Springer Berlin Heidelberg, 2009.403-416.

[8] Bai, Shuyong, Linlin Li, and Chew Lim Tan. "Keyword spotting in document images through word shape coding." Document Analysis and Recognition, 2009.ICDAR'09.10th International Conference on IEEE, 2009.

[9] Tarafdar, Arundhati, et al. "Shape code based word-image matching for retrieval of Indian multi-lingual documents." Pattern Recognition (ICPR), 2010 20th International Conference on IEEE, 2010.

[10] Wshah, Safwan, Gaurav Kumar, and Venu Govindaraju. "Script independent word spotting in offline handwritten documents based on hidden markov models". Frontiers in Handwriting Recognition (ICFHR), 2012 International Conference on IEEE, 2012.

[11] Frinken, Volkmar, et al. "A novel word spotting method based on recurrent neural networks." Pattern Analysis and Machine Intelligence, IEEE Transactions on 34.2 (2012): 211-224.

Authors

Blessy Varghese is currently a graduate student pursuing Masters in Computer Engineering at PIIT, New Panvel, and University of Mumbai, India. She has received her B.E in Computer Engineering from University of Mumbai. She has 2 years of past experience of teaching. Her areas of interest are Information retrieval, Image Processing and Natural Language processing.

Sharvari Govilkar is Associate professor in Computer Engineering Department, at PIIT, New Panvel, and University of Mumbai, India. She has received her M.E in Computer Engineering from University of Mumbai. Currently she is pursuing her PhD in Information Technology from University of Mumbai. She is having 17 years of experience in teaching. Her areas of interest are text mining, Natural language processing, Compiler Design & Information retrieval.