International Journal of Engineering Applied Sciences and Technology, 2016
Vol. 2, Issue 2, ISSN No. 2455-2143, Pages 6-10
Published Online December-January 2016 in IJEAST (http://www.ijeast.com)

A SURVEY ON WORD SPOTTING TECHNIQUES FOR DOCUMENT IMAGE RETRIEVAL

Dr. S. Vijayarani
Assistant Professor, Department of Computer Science, Bharathiar University, Coimbatore

Ms. A. SAKILA
Research Scholar, Department of Computer Science, Bharathiar University, Coimbatore

Abstract - Document images are becoming more popular in today's world and being made available over the internet, scanned/captured documents are used in paperless offices and digital libraries. Paper documents can be converted into digital form by using digitization equipments and it is stored in document image databases. If the documents are stored in image formats, it is very difficult to perform the searching process. Conventional way of information retrieval from document images into their text formats is done by using Optical Character Recognition (OCR). OCR is a technique which converts the document images into text and then we can perform the information retrieval. But the drawback here is OCR fails to perform 100% accurate conversions, hence it is difficult to perform the information retrieval task in document images. For this reason, there is a need for inventive method to search keywords in document images without performing the conversion process. Word spotting technique is one of such technique which can perform the task in two ways, i.e. segmentation based and segmentation free based methods. The steps required for these methods are discussed in this paper.

Keywords - ***Document Images, Information retrieval, OCR, Word spotting techniques***

I. INTRODUCTION

Traditional libraries and organizations are still handling large amount of hard copy/printed materials that are expensive and bulky, hence each and every libraries and offices are slowly getting digitized [1]. Paper documents can be converted into digital form by using digitization equipment like scanners, digital cameras and mobile phones (smart phones, iPhone, iPods). The most common format for these historical printed documents is the text in which the characters of the documents are represented by the machine-readable codes (e.g. ASCII codes) [3]. Extracting information from the document images is challenging problem as it compared with digital texts [17].

Information retrieval from document images has become a growing and challenging problem. Recognition and extraction of text in document images is the aim of document image analysis. However information retrieval is concerned with content based document browsing, indexing and searching from a huge database of document images. The text retrieval from document images has made significant progress and addressing related information processing problems such as topic clustering and information filtering. Information retrieval (IR) form document images are developed using two techniques.

The first approach is called recognition based retrieval which is based on Optical Character Recognition (OCR). OCR is a technique which is used to identify the characters from document images, then converts these images into their text format [8]. After conversion, documents can be edited, searched and stored [8]. This technique provides a full alphanumeric recognition of the characters in the document images [5] [6] [8]. Many different types of OCR tools are available today, but only few of them are open source and free [8]. Achieving 100% accuracy result is not possible, but it is better to have something rather than nothing [7] [8]. To improve accuracy most of the OCR tools use dictionaries, recognizing individual characters then it tries to recognize entire words that exist in the selected dictionary [8]. Sometimes it is very difficult to extract text because different font size, style, symbols, dark background and poor quality of document image often prohibit complete conversion using OCR [8]. Hence, there is a need for an alternate approach is required for finding keywords from of document images.

The second approach is recognition free retrieval which is based on Word spotting technique and it is used to match and retrieve information from document images without any conversion. It finds the user specified keyword from document image by a word-to-word matching [16, 17, 18]. A beginning study conducted in this direction with some prior results on searching word images using the word spotting technique. The system

accepts textual query from users, the query is first converted to image (template image) by rendering. For finding keywords, similarity measures are calculated and distance between relevant words are matched in the document images. This survey discusses the word spotting techniques and its methods.

The remaining portion of this paper is organized as follows. Section II presents the review of literature. Section III provides detailed description of word spotting techniques and its methods. Conclusion is given in Section IV.

II. RELATED WORK

B. Gatos, et.al [14] proposed a method which is based on block-based document image descriptors that are used at a template matching process satisfying invariance in terms of translation, rotation and scaling. Improvement in terms of time expense is obtained by applying the matching process only on salient regions of the image.

Nikos Vassilopoulos, et.al [12] analyzed classification-free Word-Spotting system, appropriate for the retrieval of printed historical document images. Moreover, it does not include segmentation, feature extraction and clustering or classification stages. Instead it treated the queries as compact shapes and used image processing techniques in order to localize a query in the document images.

Sayantan Sarkar [21] proposed the technique of word spotting using Modified Character Shape Code to Handwritten English document images. It is different from other Word Spotting techniques as it has implemented two level of selection for word segments to match search query. First one is based on word size and the next is based on character shape code of query.

Yue Lu, **et.al [1]** has proposed a method which is used to find a word portion in document images, to facilitate the detection and location of the user-specified key words. Each word image extracted from documents is represented by a feature string. An inexact string matching is used to measure the similarity between the two feature strings, based on document word image and it is relevant to the user-specified word and resolves whether its portion is the same as the user-specified word and display the result on document image.

III. WORD SPOTTING TECHNIQUES

Many numbers of techniques are involved to perform information retrieval task in document images [18]. Recognition free Information retrieval is based on Word Spotting technique and it is used for searching keywords from document images, it searches for most related keyword from image as per user request by using only image features [9]. Two different types of methods are used to perform the word spotting technique they are Segmentation based and Segmentation free method.

1. SEGMENTATION BASED METHOD

This type of word spotting technique performed its task by using segmentation. This technique involves the following steps for searching keywords from document image. They are preprocessing, edge detection, segmentation, feature extraction and matching.

1.1 Preprocessing

The document image is the raw input for document analysis. It can be digitized by using optical scanning, digital camera and mobile phone, that yield a file of picture elements or pixels, but this document image may contain noise. Hence to improve quality and reduce the noise in the images we can use preprocessing phase and it includes binarization and noise reduction.

1.1.1 Binarization

To convert the color image into a binary image or to convert an image from color image to black and white image; it is called as binary image. This method is based on various color transforms. According to the R, G, B value in the image, it calculates the gray scale values and obtains the gray image at the same time [20]. Matching technique can be easily performed on grey images or edge images.

1.1.2 Noise Reduction

Images are taken with both digital cameras, scanned images and conventional film cameras will pick up noise from a variety of sources. Further use of these images will often require that the noise be removed. In order to perform this, many filters such as average filter, adaptive filter, wiener filter, mean filter, gaussian filter and median filter, etc are applied using de-noising. These methods usually to reduce the effect of noise on the performance of the document image analysis.

1.2 Segmentation

Segmentation can be considered as the key issue in document images and it has two levels; line segmentation and word segmentation. Each line in the document images may not be perfectly horizontal, they will not have so much of skew such that there is no inter line gap. Hence the lines are aligning horizontally using line segmentation. A segmented line is selected for word segmentation, it separates the text region into lines and then finally to words. As the font size of that line is not identified, the height of the line is measured which matches to its font size [21].

1.3 Feature Extraction

Feature extraction includes extracting the meaningful information from the document images. Features are extracted only once and they are saved in the database [15]. This technique extracted every word from the image file, which are capable of capturing the word similarities and discarding slight differences due to noise or font styles and size [22]. When the input data to an algorithm is too large for processing and it is assumed to be very redundant, the input data will be transformed into a reduced representation of a set of features (feature vectors). Transformation of the input data into the set of features is called feature extraction [21]. This method is useful for enormous-size images and reduced feature representation is required to complete the tasks at a fast rate such as image matching and retrieval. Word profiles, moment based features and structural features [9] are commonly used feature detection methods. Feature extraction and similarity matching the techniques used are; feature-based vectors, word shape codes, neural networks and Hidden Markov Model (HMM) [2].

1.4 Matching word images

Matching in document images can identify the query images of the documents that are most related to the query word through the extracted feature vectors [20]. This technique also called word spotting technique. Different types of matching algorithms are available for word spotting technique; they are Euclidean distance, Cosine Similarity, Normalization Cross Correlation (NCC) and Dynamic Time Warping (DTW). Euclidean distance as the measure for finding similarity, the search is usually based on similarity rather than on exact matches [25]. Compared between two word images without aligning their feature values measured to using Euclidean distance matching algorithm [18]. Cosine similarity measures the similarity between word images without aligning their feature vector values. This feature matching technique computes the cosine angle between the compared word images feature.

Normalized Cross Correlation or Dynamic Time Warping (DTW) based word image matching algorithms are most commonly used in document images to find exact matches to the query word [9]. The normalized cross correlation function computes sequences of feature vectors of two word images by aligning them against each other, then normalizing the result by standard deviation of their features. Optimal matching score gives a dissimilarity measure for the two aligned images. Most of the word spotting techniques followed the classical approach of feature vectors for feature extraction using Dynamic Time Warping (DTW) matching algorithm. DTW is a dynamic programming based matching procedure; it is a powerful matching technique to handle word variations by aligning and comparing word images with different font style, sizes and different variations, However, DTW measures the dissimilarity between word images[18]. It is used to align and compare the sets of features which have been extracted for word image retrieval [2]. The major task of matching algorithm is to compare query features with the indexed features of the word images that present in the database of documents [24].

2. SEGMENTATION FREE METHOD

In word spotting technique without performing segmentation is called as segmentation free method. Segmentation free word spotting technique has the following steps for searching keywords from document image. They are Preprocessing, Normalization, Salient region detection, block based feature extraction, Candidate image areas, Detection of word instances and Remove Overlapping result.

2.1 Image normalization

Image normalization is used for both document image and input query using a common average character height equal to feature vector representation. Feature vectors of two word images are aligned against each other, then normalizing the result by calculating the standard deviation of their features [14].

2.2 Salient region detection

This step contains an efficient word spotting procedure; it is desired to constrain the applied matching process only on certain regions of interest. These regions should correspond to the text regions of input query. A salient region is calculated using horizontal RLSA with threshold values, which corresponds to a rough text line estimation outcome [14].

2.3 Block based feature extraction

Normalized document image was split into several word instances, in order to capture transformation variations in terms of rotation and scaling. This block based feature extraction calculates different set of feature vectors based on calculating 5x5 non-overlapping pixel densities and applying word image translation [14].

2.4 Candidate image area detection

Candidate image area detection method involves matching the input query (keyword) into a document image. This is used to find the keyword correspondences on the document image that will serve as indicators of candidate image area. Each pixels of the keyword matches the most parallel words on the document image. Each set of corresponding matching result defines a candidate image area on document image. Most related result of input query, define a bounding box around the keyword on the document image [13].

International Journal of Engineering Applied Sciences and Technology, 2016
Vol. 2, Issue 2, ISSN No. 2455-2143, Pages 6-10
Published Online December-January 2016 in IJEAST (http://www.ijeast.com)

2.5 Detection of word instances

Candidate image area detection matches query input with entire document image, but cannot guarantee that they contain the query word under consideration. For this reason each pixels of the keyword finds the most similar pixels of the candidate image area. The bounding box of the keyword is transformed into the document image resulting to the gray-shaded area. This process is applied to all candidate image areas aiming to produce a set of bounding boxes that are afterwards ranked according to their matching efficiency [13].

2.6 Remove overlapping result

The candidate image areas are created using the point correspondences between the keyword pixels and the pixels of the document image. There are cases where more than one candidate image areas correspond to the same word in the document image. Each of the overlapping bounding boxes has different ranking values. On the document image, two bounding boxes are considered as overlapping since their intersection over union ratio exceeds the threshold. The bounding box that has the larger ranking value among the overlapping bounding boxes is the one kept while the others are discarded from the list. A smaller ranking value of the bounding box is discarded from the list of bounding boxes. The remaining bounding boxes are further filtered out using the following criterion that compares the aspect ratio of the bounding box to the aspect ratio of the query image [25].

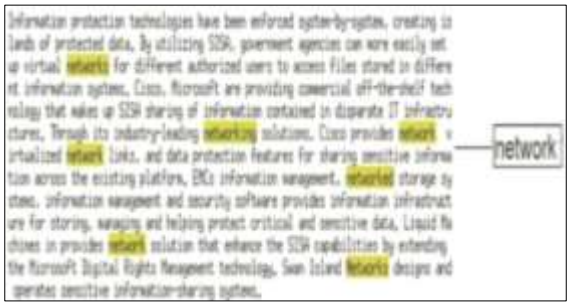

Figure1 Sample printed document being searched for the word "**network**" [9] [2]

IV. CONCLUSION

As enormous amount of document images are available in the digital libraries and other organizations, there is a need for searching strategies to find specific information from these images. Information retrieval from the document images are one of the important and challenging research problem in the field of image mining. Conventional way of information retrieval from scanned document images into their text formats is done by using Optical Character Recognition (OCR). Word spotting is an alternative approach of OCR. Two different types of word spotting based approaches for document image retrieval was explored in this paper.

V. REFERENCE

[1] Yue Lu and Chew Lim Tan "Word Searching in Document Images UsingWord Portion Matching" *Springer-Verlag Berlin Heidelberg*, DAS 2002, LNCS 2423, pp. 319–328, 2002.

[2] Blessy Varghese, SharvariGovilkar "A Survey on Various Word Spotting Techniquesfor Content Based Document Image Retrieval" *International Journal of Computer Science and Information Technologies (IJCSIT)*, ISSN: 0975-9646 Vol. 6 (3), 2015, pp. 2682-2686.

[3] Toni M. Rath, R. Manmatha "Word Image Matching Using Dynamic Time Warping"

[4] M. Mitra, B.B. Chaudhuri "Information Retrieval from Documents: A Survey" Kluwer Academic Publishers, Information Retrieval 2, 141–163 (2000)

[5] Pranob K Charles, V.Harish, M.Swathi, CH. Deepthi "A Review on the Various Techniques used for Optical Character Recognition", *International Journal of Engineering Research and Applications (IJERA)*, ISSN: 2248-9622, Vol. 2, Issue 1,Jan-Feb 2012.

[6] Shalin A. Chopra , Amit A. Ghadge , Onkar A. Padwal , Karan S. Punjabi , Prof. Gandhali S. Gurjar "Optical Character Recognition", *International Journal of Advanced Research in Computer and Communication Engineering*, ISSN (Online) : 2278-1021 ISSN (Print) : 2319-5940, Vol. 3, Issue 1, January 2014

[7] ShivaniDhiman, A.J Singh, "TesseractVsGocr A Comparative Study", *International Journal of Recent Technology and Engineering (IJRTE)*, ISSN: 2277-3878, Volume-2, Issue-4

[8] S.Vijayarani, A.Sakila "Performance Comparison of OCR Tools", *International Journal of UbiComp (IJU)*, Vol.6, No.3, July 2015

[9] Million, Meshesha and C. V. Jawahar "Matching word images for content-based retrieval from printed document images", *International Journal of Document Analysis and Recognition (IJDAR)*, DOI 10.1007/s10032-008-0067-3, Springer-Verlag July 2008

[10] Balasubramanian, A.Meshesha and M.Jawahar,C.V.: Retrieval from document image collections. In: Proceedings of the Seventh *International Association for Pattern Recognition (IAPR)*, Workshop on Document Analysis Systems (DAS), pp. 1–12 (2006)

[11] AnuragBhardwaj, SrirangarajSetlurand Venu and Govindaraju "Keyword Spotting Techniques for Sanskrit Documents"

[12] NikosVassilopoulos and Ergina Kavallieratou "A Classification-free Word-Spotting System" SPIE-IS&T, Vol.8658 86580F, pp:1-10.

[13] Thomas Konidaris, "A segmentation-free word spotting method for historical printed documents" *springer*, 17 April 2015

[14] B. Gatos, "Segmentation-free Word Spotting in Historical Printed Documents", *10th International Conference on Document Analysis and Recognition*, 2009.

[15] Pavan Kumar M. N. S. S. K. and Jawahar C. V. (2004) "Information Processing from Document Images" Information Technology: Principles and Applications (ed) Ray A. K. and T. Acharya, Prentice Hall of India, New Delhi. pp. 522—547

[16] Jeong C.B. and Kim S.H, "A Document Image Preprocessing System for Key Word Spotting" *Lecture Notes in Computer Science-2004*, Vol. 3334, pp. 440-443.

[17] Simone Marinai, " A Survey of Document Image Retrieval in Digital Libraries", *9th colloque International Francophone Sur l'Ecritet le Document (CIFED)-2006*, pp. 193–198.

[18] AdaneLetta "Feature extraction and Matching in Amharic Document Image Collections" June 2011.

[19] RangacharKasturi, Lawrence O'gorman and VenuGovindaraju (2002) "Document Image Analysis": A primer. Sadhana, 27(1): 3–22.

[20] Mr. G.T. Sutar, Prof. Mr. A.V. Shah, "Number Plate Recognition Using an ImprovedSegmentation" *International Journal of Innovative Research in Science, Engineering and Technology*, Vol. 3, Issue 5, May 2014, pp.12360- 12368

[21] Sayantan Sarkar "Word Spotting in Cursive Handwritten Documents using Modified Character Shape Codes".

[22] KonstantinosZagoris, KavallieratouErgina and Nikos Papamarkos "A Document Image Retrieval System" *Engineering Applications of Artificial Intelligence* (2010), Vol. 23, Issue 6, pp. 872-879.

[23] Manesh B. Kokare and Shirdhonkar M. S. (2010) Document Image Retrieval: An Overview. *International Journal of Computer Application*. 1(7):114-119.

[24] Chadha, Aman, SushmitMallik, and RavdeepJohar. "Comparative study of feature-extraction techniques for content based image retrieval", *International Journal of Computer Applications*, (0975–8887) Volume (2012).

[25] Reza Tavoli "Classification and Evaluation of Document Image Retrieval System" WSEAS TRANSACTIONS on COMPUTERS, E-ISSN: 2224-2872, Issue 10, Volume 11, October 2012 p.p 329-338.