ISSN:0975-9646

A Survey on Various Word Spotting Techniques for Content Based Document Image Retrieval

Blessy Varghese#1, Sharvari Govilkar*2

#1,2Department of Computer Engineering,
University of Mumbai,
PIIT, New Panvel, India

***Abstract*— Searching documents for information and retrieval of relevant documents is a basic activity. Various tools are readily available for searching and retrieval from digital documents, but not much robust methods are available for retrieval from historic documents and old manuscripts as they are not digitized but available in scanned formats. Conventional way of retrieval from scanned document images is using OCR but not all scripts can be OCRed robustly. An inventive approach to search these document images and retrieve relevant results is using keyword spotting or simply word spotting which gives better results as well is less complex than OCR. This paper discusses the different available word spotting techniques for document images.**

***Keywords*— Word spotting, document image retrieval, keyword spotting, searching and retrieval, word image.**

I. Introduction

Over the years, various ways have been studied to query on document images using the physical layout, logical structural information and other extracted contents such as image features. However, for document images containing basically text, the classical Information Retrieval (IR) approach using keywords is still often used. For such document images, conventional document image processing techniques can be easily utilized for this purpose.

For instance, many document image retrieval systems first convert the document images into their text format using Optical Character Recognition (OCR) techniques and then apply text IR strategies over the converted text documents. Several commercial systems have been developed based on this model, by first segmenting the page, then using layout analysis techniques, and then applying OCR. 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. Moreover, non-textual parts in a document image cannot be easily represented into a converted form with acceptable accuracy. In this context, it may be advantageous to explore techniques for direct characterization and manipulation of image features in order to retrieve document images containing textual and other non-textual components. 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 different relationships, and then convert them to machine-readable textual information using OCR.

On the other hand, a document image retrieval system asks whether an imaged document contains particular words which are of interest to the user, ignoring other unrelated words. Thus, essentially a document image retrieval system answers ”yes” or ”no” to the user’s query, instead of exact explicit recognition of characters and words as in the case of document information processing. 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 character, word or even document level.

The paper presents a survey of various techniques for word spotting from document images. Basic word spotting system is explained in section 2. Related work done using several techniques and past literature is discussed in section 3. Various techniques along with their features are discussed in detail in section 4 and comparison based on different criteria is discussed in section 5. Finally, the last section 6 concludes the paper.

II. Literature Survey

In this section we cite the relevant past literature that utilizes the various techniques for content based document image retrieval. Recent researches focus on word spotting techniques rather than OCR techniques for document image retrieval.

Manmatha[7] introduced the idea to segment the document into words, and then the word images are matched against each other to create equivalence classes for which the user can provide the ASCII equivalents. Two different algorithms were proposed for matching word images.

Gatos[2] presents a segmentation free approach for word spotting. As per the proposed method words are not individually segmented from a document, instead salient regions (particularly a line) are detected, document image descriptors are assigned to block region and matching is performed only on the regions of interest.

Safwan[3] 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.

Frinken[4] presents a keyword spotting method based on BLSTM neural network and then applying 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 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. The algorithm basically 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.

Serrano[5] presents a model-based approach where each sequence is first mapped to a semi-continuous HMM (SC-HMM) and an improved version of DTW is used for matching. This approach models vector sequences with a semi-continuous HMM (SC-HMM). The SC-HMM Gaussians are constrained to belong to a word-independent shared pool of Gaussians (universal Gaussian Mixture Models) as they can be learned offline and as all the states of the SC-HMMs share the same set of Gaussians, only the mixture weights contain sequence-specific information which results in a huge reduction of the computational cost. DTW algorithm is slightly modified to calculate the similarity between the states.

Fischer[6] presents an extension to the existing HMM-based spotting method by including character n-gram language models. The author points out that characters do not appear arbitrarily in any language. Motivated by this observation the author proposes statistical character n-gram models to be integrated into the spotting system for additional language context.

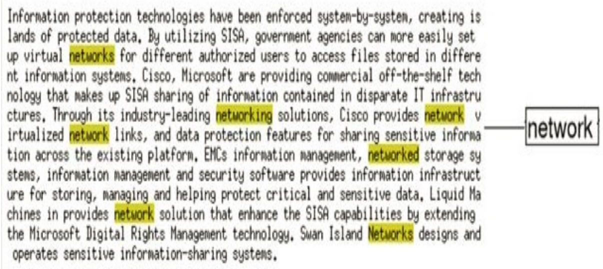
Most of the word spotting techniques uses the traditional approach of feature vectors for feature extraction followed by DTW matching algorithm. Number of different techniques for feature extraction like feature-based vectors, word shape codes, etc. and Hidden Markov Model (HMM), neural networks, etc. for similarity matching are explored by researchers.

III. Word Spotting

With storage becoming cheaper and imaging devices becoming more and more popular, attempts are on the way to digitize and archive large quantity of multimedia data (text, audio, image, video, etc). Extensive research is being carried out to make the digital content accessible to users through indexing and retrieval of relevant documents from such collection of images and text, video and audio.

Most digital libraries aim at archiving books that are not available online. Success of image retrieval systems for text mainly depends on the performance of optical character recognition (OCR), which convert scanned document images into texts. For indigenous scripts of countries like India, Ethiopia, etc., there are no robust OCRs that can successfully recognize printed text images of varying quality, size, style and font. Hence, we need an alternate approach for effective access to the large collections of document images. An improved solution is to search for relevant documents using only image properties, without explicit recognition.

This is accomplished by searching word images using the word spotting approach. The system accepts a text query from its users. This textual query is first converted to an image by rendering, by setting font, style and point size. Features are then extracted from the image and searched, for retrieval of relevant documents. Results of the search are a set of document images that are sorted in accordance to their relevance to the query word.



Information protection technologies have been enforced system-by-system, creating is
lands of protected data. By utilizing SISA, government agencies can more easily set
up virtual networks for different authorized users to access files stored in differe
nt information systems. Cisco, Microsoft are providing commercial off-the-shelf tech
nology that makes up SISA sharing of information contained in disparate IT infrastru
ctures. Through its industry-leading networking solutions, Cisco provides network v
irtualized network links, and data protection features for sharing sensitive informa
tion across the existing platform. EMCs information management, networked storage sy
stems, information management and security software provides information infrastruct
ure for storing, managing and helping protect critical and sensitive data. Liquid Ma
chines in provides network solution that enhance the SISA capabilities by extending
the Microsoft Digital Rights Management technology. Swan Island Networks designs and
operates sensitive information-sharing systems.

Figure 1. Sample printed document being searched for the word “network”[8].

A. *Phases in Word Spotting*

The input document image goes through a pre-processing phase which consists mainly of document image binarization, skew and warping correction, border detection, image normalization (common average character height), etc. After the pre-processing phase, features are detected and these features are represented using appropriate vectors or models which are later on matched for finding the similarity measure.

1) Feature extraction: It is a type of dimensionality reduction that efficiently represents interesting parts of an image as a compact feature vector. When the input data to an algorithm is too large to be processed and it is suspected to be very redundant, the input data will be transformed into a reduced representation of a set of features (also named feature vector). Transformation of the input data into the set of features is called feature extraction [12]. This approach is useful for large-sized images and a reduced feature representation is required to quickly complete tasks such as image matching and retrieval.

Various features are explored by researchers that can be considered for document images. Word profiles, moment based features, [8] etc. are commonly used for feature detection. These features are represented most commonly using vectors or sequences. Though novel approaches have started using Hidden Markov Model(HMM) and more recent approaches include different types of neural networks[4] and improvements to HMM [3][5][6].

2) Matching algorithm: Conventional approach uses Euclidean distance[9] as the measure for finding similarity. If the distance between feature vectors of the query image and images in the database is small, the corresponding image in the database is considered to be a match to the given query. The search is usually based on similarity rather than on exact matches, and then the retrieval results are ranked accordingly to a similarity index.

Correlation or DTW-based[8] word image matching algorithms are most commonly used in document images to find exact matches to the query word which restricts searching for variants of a query word in the document collections and retrieve relevant documents in the presence of writing variations in the handwritten text.

B. Word Spotting Techniques

In general, a Keyword Spotting system seeks to answer is whether a document image contains words which are of interest to the user, while paying no attention to unrelated words. In other words, a keyword spotting system provides an answer of “yes” or “no” with respect to the user’s query by spotting the keywords, rather than the exact recognition of a character as in OCR.

Moreover, words, rather than characters, are the basic units of meaning in information retrieval. Therefore, directly understanding and matching images at word level in a document image is a powerful way to retrieve information from the documents.

Keyword spotting approaches can be classified into Word spotting using Shape codes, Word spotting using Hidden Markov Model, Word Spotting using coarse features. Shape codes are generated for whole words that are being searched for; document image is trained to be represented using separate HMMs for retrieval and Word Spotting using Structural features uses features like projection profiles, moments based features, etc. Recent research works also concentrate on word spotting using different types of neural networks also.

1) Word spotting using Shape codes: Word images are represented using word shape codes [13]. These word shape codes are generated using a set of features like ascender/descender, character hole, character water reservoir, etc. Word level representation diminishes the possibility of character segmentation error as in case of OCR. Codes are generated for alphabets and numbers of English language depending on the features mentioned above. Shape codes are constructed for the query word and matched for similarity with the codes generated from the archived documents. The authors claim to achieve better retrieval results for retrieval based on the new three features presented by them although whole words can only be matched.

Researches in [14] indicate partial matching and whole word matching from the dataset. Features are extracted by using run length connected components and sandwiched vertical black and white runs in a single pass. The features used are character ascenders, descenders, deep eastward and westward concavity, holes, i-dot connectors and horizontal-line intersection [14]. Codes for 52 Roman letters are generated and each word image is defined by a concatenating these feature codes. A sequence alignment similarity measure based on dynamic programming is used for keyword spotting which, when slightly modified produces excellent retrieval results including partial matches also.

Work done in [15] allows word spotting from cursive handwritten documents by using modified word shape codes. The author, instead of segmenting the individual characters from a word, segments the word into regions based on ascender and descender, and codes are assigned to these regions which are then combined together to form the word shape token. Matching is then performed to give the final result. Modified word shape code was experimented for English language and satisfactory results are achieved by author with the limitation of more no. of false positives if query words contain less than 4 letters.

2) Word Spotting using Hidden Markov Model: Features extracted are represented using HMMs. Separate HMM can be generated for whole words or individual character HMMs generated are combined together to form HMM for a word.

Keywords are simulated as a sequence of character models, and filler models are used for better representation of background or non-keyword text. Candidate keywords are further pruned using the character based and lexicon based background models [3]. The lines in the document are segmented, skew corrected, and pre-processed to compute the feature vector by vertically dividing the line image into “bins” to compute the gradient features and these bins are further divided horizontally to calculate the intensity features. A 14-state HMM [3] with linear topology is constructed for each character. Individual character HMMs are combined together to form HMM for a word.

Viterbi beam search decoder is used to parse the entire line to find the maximum probability path between the keywords and filler models. Non-keywords are modelled using filler models which allow proper separation of keywords from non-keywords. The score of each candidate keyword is normalized with respect to the score of word background models to improve accuracy. Normalization of score is carried as a rejection strategy which reduces false positive rates resulting in better accuracy. This algorithm gave enhanced accuracy for data sets of three different languages, namely, English, Arabic and Devanagari (Hindi & Marathi)[3].

Further improvement is given in [6] which attempts to integrate character n-gram language models into the spotting system so as to provide an additional language context as the author argues that characters do not appear arbitrarily in any language. So incorporating the language context provides better spotting performance. This was experimented for English language and the data set consisted of writings from different writers.

A priori information can also be incorporated in the vector sequences for model based similarity which can further improve the retrieval results. This is discussed in [5], which models vector sequences with a semi-continuous HMM (SC-HMM) Gaussians which are constrained to belong to a word-independent shared pool of Gaussians. An interesting property of the SC-HMM is that the Gaussian components of the emission probabilities are shared and thus, can be trained separately on an unlabelled set of samples.

This allows injecting prior information about the specific domain into the model. A HMM is described by three parameters: initial occupancy probabilities, transition probabilities and emission probabilities. The proposed similarity measure is more robust than the conventional measures and this robustness is demonstrated by applying it to data sets of different languages of French, English and Arabic which gives improved results [5].

3) Word Spotting using coarse features of word-image: Coarse features of word image like the stroke density, concentration of black pixels, upper profile, lower profile, etc. are used to form the feature vectors which are later on used by the matching algorithms. Chinese word spotting [17] uses Euclidean distance as the measure to calculate stroke density feature to construct the feature vector. Modified Hausdorff distance is used for image matching.

Various features for word spotting in explored in [18] for historical manuscripts which are later on matched using the dynamic time warping algorithm as the measure for similarity. Features explored include projection profile, word profiles, background to ink transitions, grayscale variance, Gaussian smoothing, Gaussian derivatives. Projection profile is the sum of pixel values in the respective image column. A derivative of projection profile is the partial projection profile which includes three profiles: above, below and between the baselines. Upper/lower word profile features is the distance from the upper/lower boundary of the word image to the closest “ink” pixel for each image column[18]. Background to ink transitions is the number of transitions from the background to an “ink” pixel (determined by threshold). Grayscale variance is the normalized variance of the grey value intensities in every image column. Gaussian features are used generate partial derivatives like edges, smooth the image, normalize the word image into a generic height. As per the authors analysis, upper word profile and Gaussian smoothing performed best as compared to others. This was experimented on handwritten English documents.

Word profile feature is further utilized in [8] along with two new features called moments and transform domain representation. The projection and transition profiles capture the distribution of ink along one of the two dimensions in a word image, upper and lower word profiles capture part of the outlining shape of a word [8]. Moment-based features are statistical moments [such as mean, standard deviation and skew] and region-based moments [such as the zeroth-order M00 and first-order M01 moments] are computed for analysing the shape of word images. Similarity matching is performed using DTW (Dynamic Time Warping) technique which predominantly gives exact-matches, but the authors have made improvements to the algorithm so as to find partial matches also. This was experimented on different scripts of English, Amharic and Hindi which gave results above 90%.

Word shape features called GSC features were extracted and similarity is measured using correlation in [16]. Line and word segmentation is performed and then Gradient, Structural and Concavity (GSC) features are extracted for three languages English, Sanskrit and Arabic. Better results were produced by printed documents as compared to handwritten documents.

4) Word Spotting using Neural Network: The bidirectional long-short term memory (BLSTM) neural network can be used as a similarity measure in which the hidden layers comprised of long short-term memory blocks that are specifically designed to address the vanishing gradient problem[19]. The network was trained and evaluated for handwritten English documents which could give better results if the training set was large enough.

The various features extracted are stored in feature vectors which are fed into the network for training which later on outputs the probabilities to return the most likely label sequence for given test sequence. BLSTM network is also adopted for Hindi language in [11].

The BLSTM neural network along with CTC token passing algorithm has proved to perform better than the conventional systems of DTW and HMM based systems for handwritten documents in English language [4]. Though these systems require training, the results outperform the conventional systems.

C. Comparison of Word Spotting Techniques

1) Based on language dependency: Some of techniques discussed are specifically designed for particular languages. Word Shape codes were designed for English language and worked for only Roman script. HMM models were trained for English, Arabic and Devanagari scripts. Coarse feature based methods were experimented with Chinese language, later on it also worked well on English, Arabic, Sanskrit, Hindi and Amharic. Neural networks were trained for English and Hindi languages.

Word code generation is dependent on the features of the respective languages. Coarse features work at image level which results in lesser dependency on languages. HMM model and neural networks were trained for limited number of scripts though they are language independent.

2) Based on computational complexity: In word shape code method, lower no. of shape codes has lower computational complexity but gives rise to ambiguity. So, more shape codes are needed which increases the computational complexity. HMM model and neural networks require training which slightly increases the computational complexity. Coarse feature based method employs different types of features of the image but still the computational complexity remains optimal.

3) Based on training: Most of the techniques require pre-training of the word images for better retrieval results. Word shape codes are pre-decided, Statistical method requires HMM and Neural networks to be trained appropriately for precise results. Coarse feature based methods also produce accurate results if provided with suitable training but is not mandatory.

The following table presents the summary of various word spotting techniques along with its advantages and disadvantages.

TABLE I
COMPARISON OF WORD SPOTTING TECHNIQUES

Technique	Advantage	Disadvantage	Work Done
Word spotting based on Word shape Codes	Simple and easy method.	More word shape codes reduces ambiguity but increases complexity.	English [13][14][15]
Word spotting based on HMM Model	Easy to spot keywords from query during run time.	Predefined keywords and training is required.	English[3][5][6] Arabic[3][5] Devanagari[3] French[5]
Word spotting based on Coarse features of word image	Adapts variations of word images in fonts and sizes.	Full partial matching is not achieved. (Addresses word form variations only at end of word)	Chinese[17] Hindi[8] Arabic[16] English[8] [16][18] Sanskrit[16] Amharic[8]
Word spotting based on Neural Network	Outperforms other methods [4].	Larger the training set, better the retrieval results.	English[4][19] Hindi[11]

IV. CONCLUSIONS

Word spotting is a crucial area of image processing domain. The various applications ascertain the significance of word spotting in various fields. Developing an efficient search mechanism in the image domain requires designing an effective word image spotting technique which includes an efficient feature extraction scheme and matching algorithm.

Word based approaches for document image retrieval was explored in this paper. Word based approaches adopt touching characters easily and analyses the shapes of the words without explicit character recognition. All of above discussed systems retrieves information from document images by spotting the keywords. Selection of a proper word spotting technique for information retrieval can be done depending on the requirements.

ACKNOWLEDGMENT

I am using this opportunity to thank all the people who contributed in some way to the work described in this paper. My deepest thanks to my project guide for giving timely inputs and giving me intellectual freedom of work. I express my thanks head of computer department and to the principal of Pillai Institute of Information Technology, New Panvel for extending his support.

REFERENCES

[1] Murugappan, Abirami, Baskaran Ramachandran, and P. Dhavachelvan. "A survey of keyword spotting techniques for printed document images." *Artificial Intelligence Review* 35.2 (2011): 119-136.

[2] Gatos, Basilios, and Ioannis Pratikakis. "Segmentation-free word spotting in historical printed documents." *Document Analysis and Recognition, 2009. ICDAR'09. 10th International Conference on*. IEEE, 2009.

[3] 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.

[4] 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.

[5] Rodríguez-Serrano, José A., and Florent Perronnin. "A model-based sequence similarity with application to handwritten word spotting." *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 34.11 (2012): 2108-2120.

[6] Fischer, Andreas, et al. "Improving hmm-based keyword spotting with character language models." *Document Analysis and Recognition (ICDAR), 2013 12th International Conference on*. IEEE, 2013.

[7] Manmatha, R., and W. B. Croft. "Word spotting: Indexing handwritten archives." *Intelligent Multimedia Information Retrieval Collection* (1997): 43-64..

[8] 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.

[9] Chadha, Aman, Sushmit Mallik, and Ravdeep Johar. "Comparative study and optimization of feature-extraction techniques for content based image retrieval."*International Journal of Computer Applications (0975–8887) Volume* (2012).

[10] Sahu, Neha, R. K. Rathy, and Indu Kashyap. "Survey and Analysis of Devnagari Character Recognition Techniques using Neural Networks."*International Journal of Computer Applications* 47.15 (2012): 13-18.

[11] Jain, Raman, et al. "BLSTM neural network based word retrieval for hindi documents." *Document Analysis and Recognition (ICDAR), 2011 International Conference on*. IEEE, 2011.

[12] http://www.mathworks.in/discovery/feature-extraction.html?nocookie=true

[13] Lu, Shijian, Linlin Li, and Chew Lim Tan. "Document image retrieval through word shape coding." *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 30.11 (2008): 1913-1918.

[14] 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.

[15] Sarkar, Sayantan. "Word Spotting in Cursive Handwritten Documents Using Modified Character Shape Codes." *Advances in Computing and Information Technology*. Springer Berlin Heidelberg, 2013. 269-278.

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

[17] Lu, Yue, and Chew Lim Tan. "Word spotting in Chinese document images without layout analysis." *Pattern Recognition, 2002. Proceedings. 16th International Conference on*. Vol. 3. IEEE, 2002.

[18] Rath, Toni M., and Raghavan Manmatha. "Features for word spotting in historical manuscripts." *Document Analysis and Recognition, 2003. Proceedings. Seventh International Conference on*. IEEE, 2003.

[19] Frinken, Volkmar, et al. "Adapting BLSTM neural network based keyword spotting trained on modern data to historical documents." *Frontiers in Handwriting Recognition (ICFHR), 2010 International Conference on*. IEEE, 2010.