Keyword Spotting in Scanned Images of Historical Handwritten Devanagri Documents

Sushma S. N.
Dept. of Studies in Computer Science, University of Mysore
Mysuru- 570006, India

Sharada B.
Dept. of Studies in Computer Science, University of Mysore
Mysuru- 570006, India

ABSTRACT
Huge quantity of information is lying quiescent in historical manuscripts. This information would go wasted if it is not stored digitally. In keyword spotting, all occurrences of a query keyword image are retrieved from scanned document images. The problem of spotting words from handwritten documents is difficult due to its huge changeability in writing styles and its large vocabulary. Existing keyword spotting approach is mainly based on statistical depiction of word image. This paper presents an efficient structural depiction of word image, where the handwritten words are represented using graph based method for historical handwritten devanagari manuscripts. Experimentation is conducted on historical handwritten Shankaracharya's documents written in Devanagari. The results were promising in terms of accuracy and efficiency.

General Terms
Graph Representation, candidate words,

Keywords
Keyword spotting, segmentation, ranking

1. INTRODUCTION
Many national libraries have started massive digitization of precious historical manuscripts. Some of the examples are George Washington's papers at the Library of Congress and Isaac Newton's papers at the University of Cambridge Library [1]. The main appliance of keyword spotting is mining of the useful information these historical manuscripts contain. Detecting similar words from handwritten documents are still a challenge due to the deviating writing styles [2]. An alternative method of an absolute transcription of documents is Keyword spotting. Keyword spotting detects all instances of a given query word image from a document. Key applications of keyword spotting comprise handwritten notes, digital libraries and historical document retrieval, which involve large volumes of documents and necessitate efficient method of document information retrieval.

For printed documents, OCR will work exactly but not for handwritten documents. The first application of keyword spotting to handwritten text was demonstrated for ancient books for medieval handwritten text in Latin and Semitic alphabets [1]. The query word image is explored in the documents are detected and ranked based on similarity score between the query word image and the candidate words images in the database [2]. Depending on how the input is specified, keyword spotting approaches can be either query-by-string or query-by-example. Also the keyword spotting can be classified into two main approaches depending upon whether the scanned document image is segmented or not.Usually, high-level features and low-level features of handwriting are used in keyword spotting. High-level features such as ascenders, descenders and loops, which are typically more concerned with the object as a whole or larger component of it, do not work very well on unconstrained handwriting documents. On the other hand, low level features such as entropy, profiles, Zernike moments, point features are low-level are less informative but more reliable [8,9] when compared to High-level features. For Indic scripts especially for handwritten Devanagari words, numerous works has done using statistical representation. In this paper, a structural representation of handwritten word images is proposed. Historical handwritten Devanagari document resources of the Oriental Research Institute [ORI] @ Mysuru is used for testing and validation of different stages. The proposed method is also tested on keyword spotting of handwritten Gurumukhi scripts. The rest of the article is organized as follows. Section 2 presents a concise literature of keyword spotting. Section 3 gives a detailed explanation of proposed methodology, covering Graph representation, computing graph edit distance and ranking. Section 4 presents the experimental results, followed by conclusions and discussion in Section 5.

2. RELATED WORK
Keyword spotting can be classified into two most important approaches segmentation based and segmentation free approaches, depending upon whether the given document image is segmented into words or not. Segmentation based approach involves two stages like line segmentation and words segmentation. Some of the drawbacks of detection of words using segmentation techniques are partial occlusion and over or under segmentation. In Segmentation free approach, the document image does not need any segmentation and it uses sliding window concept for detecting of words. Many statistical features such as column features, pixel count features, gaussian filter features, local gradient histogram, sliding window, discrete cosine transformation profile, ink transition, chain code (CC), zoning, Fourier coefficients, projection profiles, upper/lower word profiles are used [17-21]. Nowadays, capturing and modeling the structural properties of objects using Graph based representation has become a trendy approach in pattern recognition domain. However, the employment of graph representation in the handwritten word spotting application is still very rare. There are only a few attempts made in handwritten recognition research with graph representation or relative concepts up to now. From the literature survey, it is observed that the methods proposed are used to detect the words of different languages. Among all these languages work on Devanagari word is not so common to that of English and Chinese words [28]. Many graph based approaches are found in the literature for matching the word images of different languages like English, Chinese, Latin. Structural representation is not applied for spotting the handwritten Devanagari words. However using graph theory, the words would be detected easily and would naturally make some space in the field of word spotting. This motivated us to make an attempt to develop a keyword spotting system for handwritten

devanagari documents based of structural representation. Devanagari is the basic script of many languages in India, such as Hindi, Sanskrit, Nepali and Marathi. Devanagari is half syllabic in nature. Devanagari script contains 13 vowels and 36 consonants. Devanagari word has three zones such as lower, middle and upper zone. The upper zone contains the modifiers, and lower zone contains lower modifiers. The upper zone and middle zone are always separated by the header line called Shirorekha. As Hindi is the Indian national language, Devanagari has got the position of national vernacular.

3. THE PROPOSED METHOD

In this framework, given query word image is matched with all the word images in the documents and similar words are spotted in the documents. A handwritten document is scanned and stored in database.

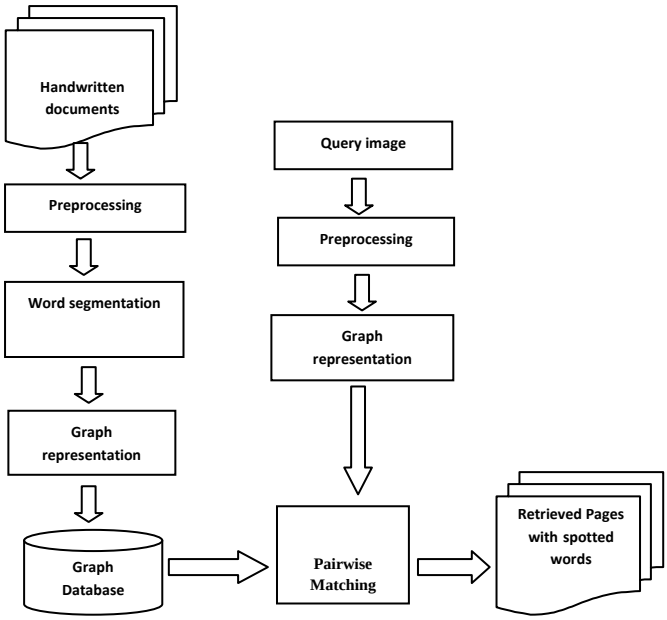


Fig 1: Overview of Keyword spotting system.

A set of operations are performed such as preprocessing, segmentation of word images [candidate words], graph representation of word images and graph extraction [using Minimum Spanning Tree]. The given query image and candidate word images are compared by pairwise matching using Graph Edit Distance [GED], corresponding similarity matching score is computed and a set of known keywords images in handwritten documents are retrieved based on the matching score. In the following subsections, detailed description of the proposed keyword spotting system is demonstrated.

3.1 Preprocessing and Segmentaion

Preprocessing operations are essential in order to improve the quality of input document images as the handwritten historical documents are characterized by low image quality and large amount of noise. In order to eliminate the noise in the image, Gaussian filtering is applied and then the image is converted into a binarized image. Based on some threshold value obtained through Otsu's method [29], skew angle correction is performed using skeletonization process. To remove all types of variations during the writing, normalization is performed and standardized data is obtained. Segmentation is performed using projection profile method. In this stage the binary image is separated into lines and words [30].

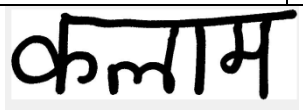
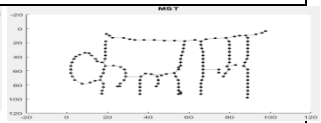
3.2 Graph Representation

In "Structural representation" for handwritten Devanagari words, each word is represented as a graph G. Let LV and LE be a finite label sets for nodes and edges, respectively. A Graph G is a four tuple (V, E, μ, ν), where,

V is set of vertices or nodes
E $\subseteq$ V × V represents set of edges
μ: V $\rightarrow$ LV is the vertex labelling function, and
ν: E $\rightarrow$ LE is the edge labelling function.

Graph extraction algorithm is based on a grid-wise segmentation of word images. Grids have been used to describe features of word images like Local Gradient Histogram (LGH) or Histogram of Oriented Gradients (HOG). Graphs are created on the basis of binarised, filtered, and skeletonised word images B. A word image B is divided into segments of equal size and for each segment a node is inserted into the resulting graph and labelled by the (x, y)-coordinates of the centre of mass. If a segment does not contain any foreground pixel, no centre of mass can be determined and thus no node is created for this segment. Finally, undirected edges (u, v) are inserted into the graph according to edge insertion algorithms, viz. Node Neighbourhood Analysis (NNA), Minimal Spanning Tree (MST).

Table 1. Minimal spanning tree representation of the segmented word images.

Segmented word Image	MST graph image
	

3.3 Graph Edit Distance (GED)

Graphs can be classified as directed and undirected graphs, where pairs of nodes are connected by directed or undirected edges respectively. Graph matching can be performed using either inexact or exact matching methods. The handwritten word matching is an Inexact graph matching method where two non-identical graphs are compared. Graph Edit Distance (GED) is an error-tolerant Inexact graph matching technique which is most powerful and flexible approach available [32]. The dissimilarity between the query word graph and all the candidate word graphs are computed using a Graph Edit Distance (GED). The main idea of the graph edit distance is that of finding the dissimilarity of two graphs by the minimum amount of distortion required to transform one graph into the other. The distortion model is composed of six types of edit operations: insertion, deletion and substitution for both nodes and edges. A sequence of edit operations (e_1, . . . ,e_K) that transforms g_1 into g_2 is called an edit path from g_1 to g_2. There are numerous edit paths available to transform g1 to g2. So, cost for each edit operation is associated. GED is taken as the least cost edit path to transform g_1 into g_2.Where, g_1=((V_1, E_1,μ_1,v_1) and g_2(V_2, E_2,μ_2,v_2) are query and candidate graphs respectively. Graph edit distance between two graphs is computed using:

$$ged(g_1, g_2) = \min_{(e1,\ldots ek)\in\gamma(g1,g2)} \sum_{i=1}^{k} c(e_i) \quad (1)$$

Where, $\gamma(g_1,g_2)$ represents the set all of edit paths to transform g_1 to g_2, c(e_i) represents the cost of the edit operation e_i.

3.4 Pairwise Matching

Mutual matching is the heart of Keyword spotting system. It is based on matching between query graph gq with the set of all word graphs G={g1,g2,…gn}in database. Sub-optimal algorithm known as Bipartite Graph Matching (BGM) [33] is used. The matching score known as rank list is computed between query graph gq with the set of word graphs in database. Ranking between gi and gq is calculated using:

$$rank(g_i) = 1 - \frac{d(g_i, g_q)}{\max[d(g_i, g_q)]} \quad (2)$$

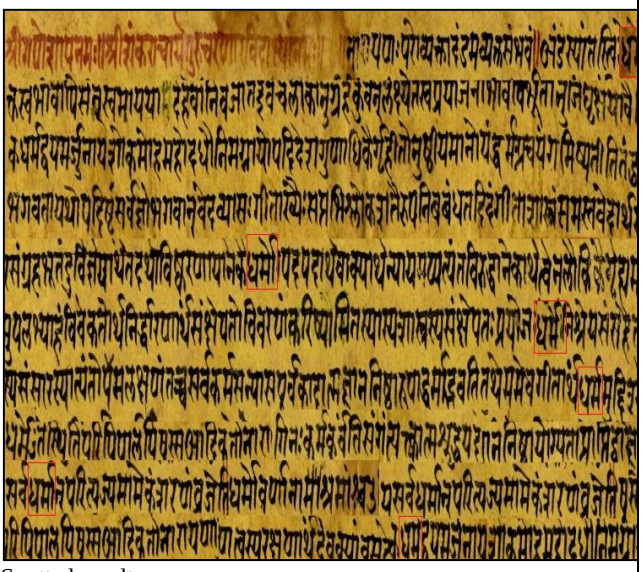
Where d(gi, gq) denotes graph edit distance between gi and gq max [d(gi, gq)] denotes maximum cost edit distance between gi and gq. Now the ranking list is normalized between 0 to 1, the ranking score is sorted and corresponding words are spotted in documents.

4. EXPERIMENTAL ANALYSIS

As per the literature review, this is the first attempt of using structural representation in keyword spotting systems for handwritten devanagari documents. At present there is no standard dataset available for handwritten devanagari document images. For the experiments two datasets of devanagari documents were used. A handwritten historical devanagari document collected from ORI @ Mysuru and a handwritten devanagari document collected from individuals of different professions. The first database contains scanned handwritten historical devanagari document images which contains 19 pages of Shankaracharya's manuscripts.

Dataset 1: Historical devanagari document

Query Keyword: Dharma 

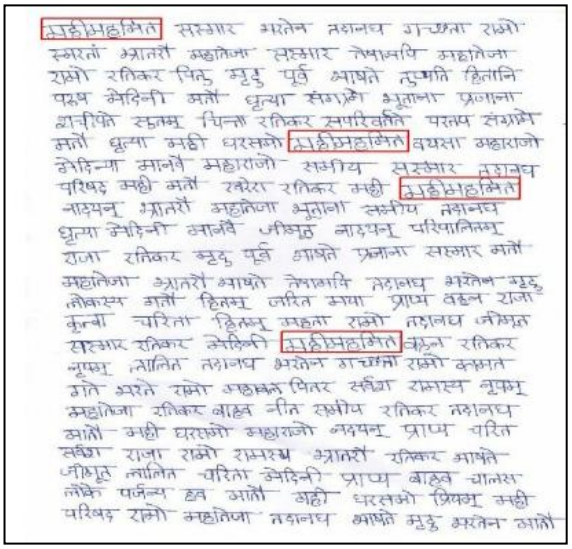


Spotted results:

Fig 2 : Qualitative results for the query "dharma" from dataset 1.

Dataset 2: Handwritten devanagari document

Query Keyword: mahemahamita 



Spotted results:

Fig 3 : Qualitative results for the query "mahemahamita" from dataset 2.

The second database contains scanned handwritten devanagari document with 23 handwritten document images. Based on the pairwise word matching criteria, similarity between query word and all candidate word images are computed and ranked. For the query word image, the top ranked candidate word images are considered and are spotted in document image.

To determine the performance of keyword spotting, Precision (PR), Recall (RC) and F-Measure (FM) are computed. These are defined as described in the subsequent section. For the evaluation of the proposed approach, 6 different query word images and its frequency of occurrences in the document image is noted. The ground truth of these words is shown in above Table. Let Wf be the total number of keyword instances and Wcs is the correctly spotted keyword instances [33]. Then,

$$Precision = \frac{W_{cs}}{W_f} \times 100 \quad (3)$$

$$Recall = \frac{W_{cs}}{W_d} \times 100 \quad (4)$$

$$F - Measure = \frac{W_{cs} \times 2(RC \times PR)}{Wf \times (RC + PR)} \quad (5)$$

Where precision is the percentage of the spotted words that exactly match the query word. Recall is the percentage of the words, same as query word that are successfully retrieved from the word database. The better performance of the keyword spotting system depends on elevated the value of F-measure. The keyword spotting performance of structural approach in terms of recall and precision as well as F-measure for devanagari documents is shown in Table 2.

Table 2. Ground truth of the sample words and its corresponding keyword spotting performance in terms of recall and precision as well as F- measure.

Sl No.	Word	Word occurrence in database	Precision	Recall	F-measure
1	dharma	24	88.42	86.19	86.94
2	maya	19	85.33	88.17	87.94
3	ratikar	16	87.71	87.5	88.17
4	mahema hamita	27	87.27	89.93	87.72

The precision and recall are represented as chart in Fig.4

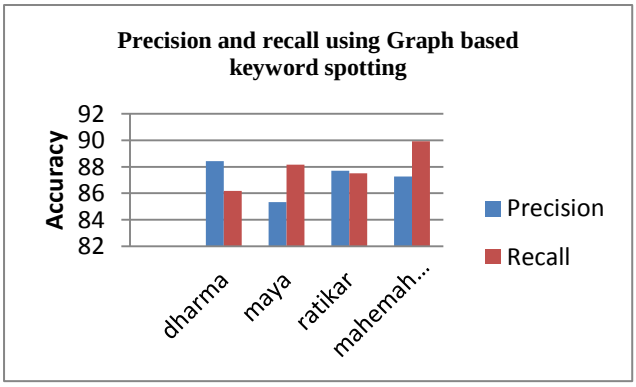


Fig 4: Precision and recall using Graph based keyword spotting

5. CONCLUSION

In this work, a graph based approach for spotting of historical handwritten Devanagari words is described. The accuracy of This work directly depends on the handwriting of the subject. Although the vast majority of word matching algorithms rely on statistical data representations, more and more attempt is now made in various research fields on structural representations. Unlike the statistical representation which ignores the dependencies between observations, graphs conserves these dependencies and relations. There are few works reported for word spotting from documents for English, Chinese, Latin and Arabic [26, 31], however, the proposed work for spotting keywords from handwritten devanagari document images using graph representation is first of its kind. In this paper, it has been shown that on a handwritten devanagari document dataset, the average keyword spotting accuracy is 87%. It can be concluded that the proposed approach effectively performs keyword spotting in handwritten divanagari documents. Finally, considering the enormous complexity of historical devanagari script, the contribution of the present approach may be considered significant with satisfactory performances. The proposed approach is well suited for keyword spotting of historical devanagari documents. It can be extended for the spotting words in various south Indian languages. The overall accuracy of the keyword spotting system depends on usage of suitable image processing technique for historical document images.

6. REFERENCES

[1] T. M. Rath and R. Manmatha, "Word spotting for historical documents," in International Journal on Document Analysis and Recognition (IJDAR), vol. 9 pp. 139–152, 2007.

[2] R. Plamondon and S. Srihari, "Online and off-line handwriting recognition: A comprehensive survey," in IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 22, pp. 63–84, 2000.

[3] T. Konidaris, B. Gatos, K. Ntzios, I. Pratikakis, S. Theodoridis, S.J. Perantonis, " Keyword guided word spotting in historical printed documents using synthetic data and user feedback," in International Journal of Document Analysis and Recognition, vol. 9, pp.167–177, 2007.

[4] J. Almazan and A. Gordo and A. Forn ´ es and E. Valveny, "Segmentation free Word Spotting with Exemplar SVMs," Pattern Recognition, 2014.

[5] J. Almazan, A. Gordo , A. Fornes and E. Valveny, "Word Spotting and Recognition with Embedded Attributes," TPAMI, 2014.

[6] S. Wshah, G. Kumar, V. Govindaraju, "Script independent word spotting in offline handwritten documents based on hidden markov models", Frontiers in Handwriting Recognition (ICFHR) 2012 International Conference on, pp. 14-19, 2012.

[7] V. Frinken, A. Fischer, R. Manrnatha, H. Bunke, "A Novel Word Spotting Method Based on Recurrent Neural Networks", IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 34, no. 2, pp. 224, Feb. 2012.

[8] I. B. Messaoud, H. Amiri, H. E. Abed, V. Margner, " Document Preprocessing System – Automatic Selection of Binarization," in IAPR International Workshop on Document Analysis Systems, 2012.

[9] T. M. Rath and R. Manmatha, "Word image matching using dynamic time warping," in CVPR, vol. 2, pp.521, 2003

[10] A. Jose, R. Serranoa and F. Perronninb, "Handwritten word-spotting using hidden Markov models and universal vocabularies," in Pattern Recognition, vol. 42, pp. 2106-2116,2009.

[11] S. Kim , S. Park, C. Jeong , J. Kim , H. Park , G. Lee " Keyword Spotting on Korean Document Images by Matching the Keyword Image, " in Digital libraries, vol. 3815, pp. 158–166, 2005

[12] C. L. Liu, J. Kim, J. H. Kim, "Model-based stroke extraction and matching for handwritten Chinese character recognition," in Pattern Recognition vol. 34, pp. 2339-2352, 2001.

[13] T. Adamek and N. O. Connor, "Efficient contour-based shape representation and matching," in ACM SIGMM International Workshop on Multimedia Information Retrieval, 2003.

[14] Y. Leydier, A.Ouji, F.L.Bourgeois and H.Emptoz, "Towards an omnilingual word retrieval system for ancient manuscripts," in Pattern Recognition vol.42, pp. 2089-2105, 2009.

[15] T. Novikova, O. Barinova, , P. Kohli, Lempitsky, " Large-lexicon attribute consistent text recognition in natural images," Computer Vision - ECCV, 2012.

[16] S. Mozaffari, K. Faez, V. Märgner and H. E. Abed ," Two-stage lexicon reduction for offline Arabic handwritten word recognition," in International Journal of Pattern Recognition and Artificial Intelligence, vol. 22, pp. 1323, 2008.

[17] A. Andreev and N. Kirov, "Some variants of Hausdorff distance for word matching," in Review of the National Center for Digitization, vol.12, pp.3–8, 2008.

[18] S. H. Cha, C. C. Tappert, S. N. Srihari, "Optimizing Binary Feature Vector Similarity Measure using Genetic Algorithm and Handwritten Character Recognition," in International Conference on Document Analysis and Recognition, vol. 02, , pp. 662, 2003.

[19] N. I. Cho and S. K. Mitra, "Warped discrete cosine transform and its application in image compression," in IEEE Transactions on Circuits and Systems for Video Technology, vol. 10, pp. 1364-1373, 2000.

[20] U. V. Marti and H. Bunke, "Using a statistical language model to improve the performance of an HMM-based cursive handwriting recognition system," in Journal of Pattern Recognition and Art. Intelligence, vol. 15, pp. 65–90, 2001.

[21] K.A.Senthildevi and E. Chandra, "Keyword spotting system for Tamil isolated words using Multidimensional MFCC and DTW algorithm," in International Conference on Communications and Signal Processing (ICCSP) , 2015.

[22] S. Abirami and D. Manjula, "Profile Based Information Retrieval from Printed Document Images," in International Conference Computer Graphics, Imaging and Visualization, vol. 3 , pp. 268-272, 2013.

[23] A. Shrivastava, T. Malisiewicz, A. Gupta, and A. A. Efros, "Datadriven visual similarity for cross-domain image matching,"in ACM TOG, vol. 30, pp. 154, 2011.

[24] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," in International journal of computer vision, vol. 60, pp. 91–110, 2004.

[25] V. Frinken, A. Fischer, R. Manmatha, H. Bunke, "A novel word spotting method based on recurrent neural networks," in IEEE Trans.Pattern Anal. Mach. Intell. Vol. 34 , pp. 211-224, 2012.

[26] V. Lavrenko, T. Rath, R. Manmatha, "Holistic Word Recognition for Handwritten Historical Documents," in *IJCA™ : www.ijcaonline.org* ibr. DIAL'04, pp. 278–287.,

[27] [M. Rusiñol, D. Aldavert, R. Toledo, and J. Llados, "Browsing Heterogeneous Document Collections by a Segmentation-free Word Spotting Method.," in ICDAR., 2011.

[28] S. N. Srihari, H. Srinivasan, C. Huang and S. Shetty, "Spotting Words in Latin, Devanagari and Arabic Scripts," in Indian Journal of Artificial Intelligence, vol.16, pp. 2-9, 2006.

[29] A. Papandreou, B. Gatos, G. Louloudis, and N. Stamatopoulos, "Document image skew estimation contest," in ICDAR, pp. 1444–1448, 2013.

[30] M. Kumar, M. K. Jindal and R. K. Sharma, "k -Nearest Neighbor Based Offline Handwritten Gurmukhi Character Recognition," in International Conference on Image Information Processing, 2011.

[31] K. Riesen., S. Emmenegger, H.Bunke, " A novel software toolkit for graph edit distance computation. In: Graph-Based Representations," in Pattern Recognition, pp. 142-151, 2013.

[32] K. Riesen, H. Bunke " Approximate graph edit distance computation by means of bipartite graph matching," in Image and Vision Computing, vol. 27, pp. 950-959, 2009.

[33] J. L. Rothfeder, S. Feng and T. M. Rath, "Using corner feature correspondences to rank word images similarity," in Computer Vision and Pattern Recognition Workshop, pp. 30-35, 2003.