Foreground and Background information based Word Spotting for Kannada Documents

Somnath Biradar[1] Veershetty C[2] Prabha[3]
[1]Department of Electronics Engineering [2,3]Department of Studies and Research in Computer Science
[1,2,3]Karnatak Arts, Science and Commerce College Bidar, India

Abstract— In this paper, we proposed a method foreground and background information based word Spotting for Kannada documentsusing morphology filters. Morphological filters employed to capture global properties of the underlying image. An input document image is segmented into words and stored into a library. Then morphology filters are employed to represent the words of a large dataset as well as a query word. Then the cosine distance is used to measure the similarity between two words, based on it,the relevance of the word is estimated by generating distance ranks. Then correctly matched words are selectedat different distance thresholds such as97, 98 and 99 percent. Encouraging results are achieved in terms of the average precision rateat 79.20, average recall rate as 91.00 and F measure as84.57asa threshold 98%.

Key words: Word Retrieval, Morphology filters, Digital Image Processing (DIP), Image Retrieval (IR)

I. INTRODUCTION

These days' computers and internet usages are common in everywhere, almost all the documents are preserving in digital format in the form of image without index and adequate information. Empowering indexing and skimming over vast transcribed databases is a tricky objective in document examination. Traditional OCR advancements are accessible successfully for non-European languages and not for Indian language like Kannada Hindi Devanagiri etc. Kannada lacks a standard test bed of character images for OCR performance evaluation. As a consequence the methodology for Kannada OCR should be font and size independent. It must also be scalable for including variety of fonts for training with little effort; therefore word spotting is an alternative to OCR. Indian connection, this procedure is the best different option for multilingual OCR. Actually, outlining of multi-script OCR for Indian scripts is an extremely complex undertaking. Consequently, word retrieval from all Indic scripts should be tried and accentuation must be given to add to a nonexclusive word spotting algorithm. Subsequently, in this paper, we have stressed on two angles, one is an extraction of hearty components other, retrieval of Kannada words for document retrieval, document classification and language identification for which still OCR does not exist. The remainder of the paper is described as follows: the literature is explained in Section 2, in Section 3 the proposed methodology is presented, in Section 4 results and discussions are presented and finally in Section 5 conclusions are drawn.

II. RELATED WORK

Word spotting is an alternative way for OCR with respect to content based image retrieval, spotting the word or retrieval similar words in a document image; here a query word is key component. The query can be either (Query-By-Example (QBE)) or a string containing the word to be searched (Query-By-String (QBS)). Methods following QBE pattern grants an enormous disadvantage in practical applications as in order to spot a word the user needs to first locate/input an instance of such word. On the other hand QBS methods allow the user to type the keyword to search in a much more natural way. The terms Word Retrieval and Word spotting [15] are utilized reciprocally in previous work referring to a collection of methods which mean to give solutions for retrieval query word in printed and written by hand document images. Initially word spotting is reported in [3] for hand written document image and later printed document in [2].

Early methods on word spotting followed a related pipeline as OCR technologies, preliminary with binarization followed by structural/layout study and segmentation at line, word and/or character level. Several works [25][11] have been stated to find keywords and retrieve document images straight without OCR. One method is through the character shape coding [4] which converts imaged characters into a set of pre-defined codes. The created character shape codes are then gathered into higher-level word shape symbols. Lastly, keywords are situated and document images are retrieved by using the determined word shape tokens.

Profiles or pixel-based features are more qualified to get word representations, which are dialect free, instead of structural feature and shape-codes, which are characterized to catch the particular states of the written symbol images of dialect .A mix of a few sorts of feature a is frequently used to catch shapes and spatial data at various levels of granularity .This is the situation for the Gradient-based binary feature (known as GSC) representations blending pixel-based elements and auxiliary components [9], the mix of profiles with factual components [19] or with pixel thickness in zones [16]. In recent works [23], authors proposed HMMs with application to writer identification forword spotting the method exploits the continuous model into shape codebook and word specific parameters separation further in [24], presented word spotting for printed historical documents based on shapefeature followed by sequence comparison, shapes feature are extracted as a scharacter types byadopting the following vertical projection calculated on gray scale image, upper and lower profiles, ink transition of every column of an image, vertical histogram and transitional status of middle row of an image. In [8], compound features plan was accounted for comprising of projection profiles, upper/lower word profiles and background to-ink transition. Word profiles, structural feature and statistical movements were applied in [10], and further, inventors utilized Fourier coefficients for dimensionality diminishment. Different separation measures were utilized as a part of evaluating the likeness amongst query and database words image as talked about in [6, 7]. Further, dynamic time wrapping (DTW) was every now and again utilized strategy to determine the likeness of the words as utilized as a part of [17, 18] because of the fact that it tolerates unique dissimilarity not at all like above technique.

III. Proposed Approach

The general approach taken by [10,11] is to segment individual words from a documents, extract a set of features from each word, and compare all the words to each other via their feature sets using the cosine similarity measure, according to the results of the cosine similarity distance with different thresold. Similar words are retrieved from documents with respect to query word.Fig-1.Shows the flow diagram of the proposed method.

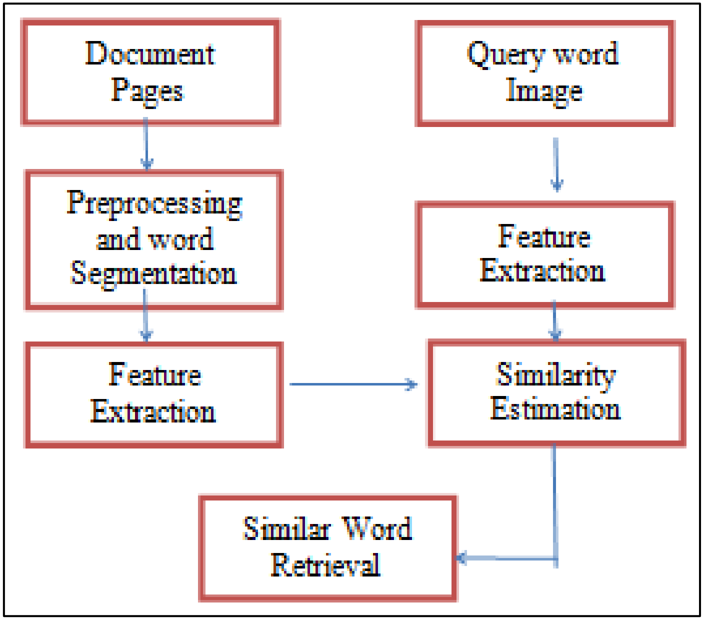

Fig. 1: The flow diagram of the proposed approach

A. *Preprocessing*

HP scanner is used for scanning the documents at 300 DPI, which usually yields a low noise and good quality document image.Thereafter Otsu's global thresholding approach is applied to convert gray scale image to two-tone images. Threshold is a normalized intensity value that lies in the range [0, 1].Otsu's method chooses the threshold to minimize the interclass variance of the threshold black and white pixels [1]. The two-tone images are then converted into 0-1 labels where the label 1 represents the object and 0 represents the background. The smaller objects like, single or double quotation marks, hyphens and periods, etc. are removed using morphological opening. After this by connected component rule each word marked as a bounding box and these bounding box words are isolated from documents image and generated 10000 words dataset.

B. *Feature Extraction*

Feature extraction is the technique for extracting the important properties of the underlying image. In this paper we computed 20 featuresset-1 for the foreground information of the image and 20 featuresset-2 background information of the same image then combined set-1(F1) and set-2(F2) to form total feature set-3(F3) set of the query word as well as database. The detail of feature extraction technique is described below

Let g(x,y) is the input image , directional opening, opening by reconstruction and top and bottom hat basic transformations applied on the input image with a structuring element. The length of structuring element (μ) is constructed based on height of the word character components with some constant K, experimentally fixed 0.7 for reconstruction and as well as for opening 0.5 to get desired directional erosion image with different orientation θ = 0, 45, 90 and 135 degrees. And also the density estimation of directional reconstructed image after hole filling is computed using equation (1). For illustration, the process of reconstruction in different direction is shown in Fig. 2.

$$\sigma_{(\theta\ \mu)} = \frac{\sum_{i=1}^{R} \ \sum_{j=1}^{C} g(i,j)}{RxC} \qquad (1)$$

On-pixels Ratio(η): A pre-processed input image g is directly used for hole filling and an output image g'$_{hf}$ is obtained. The arithmetic means of g'$_{hf}$ is defined as

$$\eta = \frac{\sum_{i=1}^{R} \ \sum_{j=1}^{C} g\prime hf(i,j}{RxC} \qquad (2)$$

Thus, 11 features were obtained. We call this method as Feature set-1 (F-1) and to make the features independent of font size; they were normalized by dividing them with the size of the image. For illustration, the process of reconstruction in horizontal direction is shownin Fig. 1

In addition to the above 11 features, we added another 9 features which were computed and normalized using (3) and (4). This set of features was also globally extracted and called this method as Feature set-2 (F-2). Two different normalization techniques were employed to design F-1 and F-2 methods.

$$\sigma^{*}= \frac{\sum_{i=1}^{R} \quad \sum_{j=1}^{C} g\prime dop \;\; g(i,j}{\sum_{i=1}^{R} \quad \sum_{j=1}^{C} g(i,j)} \qquad (3)$$

The combination of these two feature sets is termed as Feature set-3 (F-3) and it consists of 20 features

$$\eta^{*}= \frac{\sum_{i=1}^{R} \quad \sum_{j=1}^{C} g\prime hf \, g(i,j}{\sum_{i=1}^{R} \quad \sum_{j=1}^{C} g(i,j)} \qquad (4)$$

where g is the output image after hole filling, R is the number of rows and C is the number of columns of g.

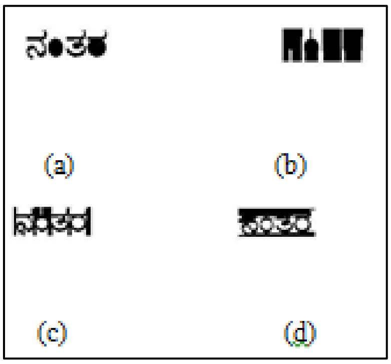

Fig. 2: Visualization of a word image after applying morphology filters (a) Reconstruction Image (b) Opened Image (c) Tophat Image in 90 degree(d) Tophat Image in 180 degree

C. Cosine Similarity

This metric is much of the time utilized when attempting to decide similitude between two document or word image, In this likeness metric, the properties (or words, on account of the documents) is utilized as a vector to determine the standardized dab result of the two words. By deciding the cosine comparability, the user is adequately attempting to discover cosine of the point between the two words. For cosine likenesses bringing about an estimation of 0, the reports don't offer any properties (or words) on the grounds that the point between the sequences is 90 degrees. Expressed as a mathematical equation:

$$Similarty(x,y) = cos\theta = \frac{x.y}{\|x\|.\|y\|} \qquad (9)$$

IV. EXPERIMENTS

A. Dataset

There is no openly accessible dataset of Kannada document images at present. Along these lines, we have gathered 50 document pages has a place with different fields, to be specific writing, Literature and history. Then the documents are scanned at 300 dpi resolution utilizing an HP scanner. A dataset of10000 words is produced from these document pages. This dataset includes expressions of various textdimension, style, length and thickness (pixel thickness).

B. Evaluation Protocol

To assess the execution of the technique, physically a ground truth is readied for 4 diverse query words which have numerous events in the dataset. The ground truth of the database is appeared in Table-1.

Sl.No.	Document Categories	No of Pages	Keyword Image	Word Frequency in database
1	Literature	14	ನಂತರ	42
2	History	26	ಬೌದ್ಧ	40
	.		ಇತಿಹಾಸ	22
			ಭಾರತ	09

Table 1: ground truth of the database

V. RESULT ANALYSIS

The eventual outcomes of word recuperation from Kannada documents are shown in Table-2. The closeness is measured between the query and candidate word image using cosine similarity. The significance of the word is measured by creating partitions positions. By then, precisely organized words are picked at different thresholds for case 97, 98 and 99 percent. To gage the execution of wordrecuperation, Recall (RC), and Precision (PR) are used. These are defined as depicted in the going with area [3][4].

$$\text{Recall} \;=\; \frac{\text{Total SamewordRetrieved}}{\text{Total SameWord Existing}} * 100 \qquad (5)$$

$$\text{Precision} \;=\; \frac{\text{Total SamewordRetrieved}}{\text{SameWord Existing+False Positives}} * 100 \; (6)$$

$$\mathrm{F-measure} = \frac{2*(\mathrm{Recall}*\mathrm{Precision})}{(\mathrm{Recall}+\mathrm{Precision})} \qquad (7)$$

We have accomplished empowering results as far as normal review, exactness and F-measure. In Table 2, we can see that the average F measure is 84.57% which demonstrates the huge execution of the technique. The execution of the proposed strategy is likewise assessed regarding Recall (RC), and Precision (PR) and they are 91.00% and 79.20% when limit (threshold) is set to 98%.

Image	Threshold ∂ −>	99.0%	98.0%	97.0%
	Recall (%)	77.50	88.57	88.57
	Precision (%)	73.80	73.80	73.50
	F-measure (%)	75.60	80.51	80.55
	Recall (%)	100.00	90.00	90.00
	Precision (%)	87.50	90.00	90.00
	F-measure (%)	93.33	90.00	90.00
	Recall (%)	100.00	100.00	76.00
	Precision (%)	81.36	86.36	86.36
	F-measure (%)	89.99	92.68	80.84
	Recall (%)	66.66	85.71	66.66
	Precision (%)	66.66	66.66	66.66
	F-measure (%)	66.66	74.99	66.66
Average	Recall (%)	86.04	**91.00**	80.39
	Precision (%)	77.47	**79.20**	79.20
	F-measure (%)	85.72	**84.57**	79.60

Table 2: Presents the word retrieval results for 4 keywords in terms of Recall and precision as well as an F- measure

VI. Conclusion

In this paper, retrieval of words from Kannada documents in light of morphological filters is proposed. This is an partial attempt of documental retrieval. The strategy has demonstrated its exceptional execution as far as Precision and Recall. Fundamentally, word spotting is vital to evade complete translation of the document image to alter/read it electronically. Especially in Indian setting, this strategy is the best distinct option for multilingual OCR. Truth be told, outlining of multi-script OCR for Indian scripts is extremelyintricate assignment. Hence, word recovery from all Indic scripts should be tried and accentuation must be given to build up a nonspecific word spotting calculation. In this paper, as an underlying endeavor word recovery from Kannada script is completed and it will be stretched out to all Indic scripts in future to build up a bland calculation.

Acknowledgement

This work ispart of the UGC Minor Research Project(RefNo:1411MRP/14-15/KAGUO13/UGC-SWR Dated4-2-2015). The prime author is thankful to UGC, New Delhi.

Reference

[1] N.Otsu,A, threshold selection method from gray-level histograms, Pattern Analysis and Machine Intelligence,vol. 9(1), pp.62–66,(1979).

[2] Shyh-shiaw, Kuo and Oscar E. Agazzi, Keyword Spotting in Poorly Printed Documents Using Pseudo 2-D Hidden Markov Models IEEE, pp.0162- 8828,(1994).

[3] R. Manmatha, C. Han, E.M. Riseman, Word Spotting, A New Approach to Indexing Handwritings. Proceedings of Computer Vision and Pattern Recognition, pp.631-637, (1996).

[4] R.Manmatha,C.Han,E.M.Riseman,WordSpotting,A New Approach to Indexing Hand writings, Proceedings of Computer Vision and Pattern Recognition,pp.631-637,(1996)

[5] B.S.Manjunath, and W.Y. Ma, Texture Features for Browsing and Retrieval of Image Data, Pattern Analysis and Machine Intelligence, vol. 18. (8), pp.837-842, (1996).

[6] Y. Lu and C.L. Tan, Word Spotting in Chinese Document Images without Layout Analysis. Proceeding of 16th International Conference on Pattern Recognition, pp. 57-60, (2002).

[7] R. Manmatha and T.M. Rath, Indexing of Handwritten Historical Documents—Recent Progress, Proceeding of Symposium on Document Image Understanding, pp. 77-85, 2003.

[8] Rath TM, Manmatha R. Word Image Matching using Dynamic Time Warping. In: Computer Vision and Pattern Recognition, pp. 521–527, 2003.

[9] Bin Zhang, S.N. Srihari, C. Huang, Word image retrieval using binary features, in: E.H.B. Smith, J. Hu, J. Allan (Eds.), Proceedings Document Recognition and Retrieval XI, SPIE, Bellingham, pp. 45–53, 2004.

[10] Jawahar CV, Balasubramanian A, Meshesha M. Word Level Access to Document image Datasets. In: Proceedings of the workshop on Computer Vision Graphics and Image Processing (WCVGIP), pp. 73–76, 2004.

[11] Lowe, D., Distinctive image features from scale-invariant key points, International Journal of Computer Vision, vol. 60, pp. 91110, 2004.

[12] Z. Shi and V. Govindaraju,, Historical document image enhancement using background light intensity normalization, in International Conference in Pattern Recognition, vol. 1, 2004.

[13] Dalal, N. and Triggs, B., Histograms of oriented gradients for human detection, in IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 886893. 2005.

[14] N. Chen and D. Blostein, A survey of document image classification: problem statement, classifier architecture and performance evaluation, International Journal of Document Analysis and Recognition , pp. 1–16, 2007.

[15] T.M. Rath, R. Manmatha, Word spotting for historical documents, Interna- tional Journal of Document Analysis and Recognition 9 ,pp. 139–152, 2007.

[16] T. Konidaris, B. Gatos, K. Ntzios, I. Pratikakis, S. Theodoridis, S.J. Peran-tonis, Keyword-guided word spotting in historical printed documents using syn- thetic data and user feedback, International Journal on Document Analysis and Recognition 9 167–177. 2007.

[17] Brina C.D, Niels R, Overvelde A, Levi G, Hulstijn W. Dynamic Time Warping: A New Method in the Study of Poor Handwriting. Human Movement Science 27(2):242–255, 2008.

[18] Forne´s A, Llado´s J, Sa´nchez G Old Handwritten Musical Symbol Classification by a Dynamic Time Warping Based Method. Graph Recognition vol 5046:51–60, 2008.

[19] M. Meshesha, C.V. Jawahar, Matching word images for content-based retrieval from printed document images, International Journal on Document Analysis and Recognition 11 (1) 29–38, 2008.

[20] Siddiqi and N. Vincent, “A set of chain code based features for writer recognition,” in International Conference on Document Analysis and Recognition , 2009, pp. 981–985, 2009.

[21] M. I. Shah and C. Y. Suen, Word spotting techniques in document analysis and retrieval-a comprehensive survey, Handbook of Pattern Recognition and Computer Vision, vol. 4, pp. 353–376, 2010.

[22] MallikarjunHangarge, B.V.Dhandra, Offline Handwritten Script Identification in Document Images, International Journal of Computer Applications (0975 – 8887) Vol. 4, pp. 6, 2010.

[23] José A. Rodríguez-Serrano, FlorentPerronnin Gemma Sánchez, JosepLladós , Unsupervised writer adaptation of whole-word HMMs with application to word-spotting, Pattern Recognition Letters 31 742–749. 2010.

[24] KhurramKhurshid ,ClaudieFaure , NicoleVincent, Word spotting in historical printed documents using shape and sequence comparisons, Pattern Recognition 45 2598–2609. 2012.