Handwritten Document Retrieval System for Tamil Language

AN. Sigappi
Department of Computer Science and Engineering
Annamalai University
Annamalainagar-608 002

S. Palanivel
Department of Computer Science and Engineering
Annamalai University
Annamalainagar-608 002

V. Ramalingam
Department of Computer Science and Engineering
Annamalai University
Annamalainagar-608 002

ABSTRACT

The paper attempts to create a handwritten document retrieval system suitable for Tamil language, with a view to record traditional literature content for future reference. It projects a search mechanism to access the query word images using a statistical model based methodology. The scheme revolves around a well defined procedure which results in word models from where the search word can be recognised and the relevant documents retrieved. The approach involves the use of hidden Markov models (HMM) to characterize the features of the dynamically varying strokes of handwritten characters. The strategy is investigated for a sample document set over a commonly used literature. The results reveal that the system yields a reasonable performance with considerable accuracy. The highlight of this procedure is that it can effectively segment differently written words from text lines in a document and imbibes in it a flexibility to cover a wide range of tilts in the strokes that are attached to the different words.

General Terms

Pattern recognition, Image segmentation, Document retrieval

Keywords

Handwritten document retrieval, Profile features, Segmentation, Hidden Markov models.

1. INTRODUCTION

Tamil is an ancient, classical Dravidian language in existence for over two thousand years. The Tamil script traces its roots to the Brahmi script and continues to undergo a lot of changes to transgress itself as a portable medium. A total of two hundred and forty seven characters constitute the Tamil language and explicitly seeks the role of rounded letters. Over the past few centuries a large collection of written and printed material is archived in many libraries across the world. The shear volume of the stored documents makes it very difficult to search and locate the required information. Though efforts are in vogue to preserve the handwritten and printed paper documents in the form of micro fiches and scanned images, which are only replicas of the physical document, still no information regarding the actual content is available. It still remains a fact that no scientific and formal methodology is available to actually search the handwritten documents for specific words.

With the advances in image processing and pattern recognition technologies, there is an emergence of a scope for searching these documents for words even in the image formats. The recognition of words in printed documents is comparatively easy as they are in standard type faces and fonts. However hand written documents present unique challenges in the sense that they are not geometrically uniform. Although research in this field over the last two decades orients to successfully address some of the issues pertaining to English and a few other non Latin script based languages, it is yet to evolve comprehensive solution strategies for indexing, searching, and retrieving handwritten documents in most regional languages, such as Tamil, Hindi, Telugu, among many others. It is in this context, that a handwritten document retrieval system is designed to accquire the documents that contain the search word from an illustrative set of handwritten documents in Tamil language.

2. RELATED WORK

There has been substantial efforts in analyzing handwritten documents for segmenting text lines and words, extracting features that characterize the word, and creating statistical models that aid in recognition of words during document searches. The text line segmentation of hand written documents in their distinct entities based on the distinction of inter- and intra-word gaps using the combination of two different distance metrics, namely Euclidean distance metric and the convex hull-based metric has been proposed [3]. The application of ANN as an aid to segmentation of handwritten characters in Assamese has been explored and reported to tackle non-touching hand written characters in addition to separating out partially touching characters [4]. The use of a segmentation free approach that performs spotting and segmentation concurrently using a sliding window has been suggested [9].

A number of features suitable for word image matching, such as word profile features, gaussian smoothing, and gaussian derivatives, have been used with the dynamic time warping algorithm for word image retrieval on a subset of George Washington's manuscript collection [11]. A probabilistic model of word spotting has been described that integrates word segmentation and word recognition probabilities, obtained by modeling the conditional distribution of multivariate distance features of word gaps and the distances returned by the word recognizer, respectively [2]. A word-image matching scheme that achieves high performance in the presence of script variability, printing variation, degradation, and word form variants has been proposed. It has been found to include a partial matching algorithm for morphological matching of word form variants in a language [5]. An approach for writer adaptation in a word spotting task has been proposed to derive writer specific word models by statistically adapting an initial universal codebook to each document, using semi-continuous hidden Markov model [6]. The design and functionality of a search engine on handwritten documents based on global image features and search using image-image similarity measure and a text-to-image score has been described [10]. A document image retrieval system that locates words in document image archives performs feature extraction followed by image matching using Minkowski L1 distance has been proposed [1].

3. PROBLEM STATEMENT

Every language imbibes a well defined format for inscribing the characters and thus creates an astute challenge to identify the most appropriate features that characterize the words in a language. The main objective is to design a system that is capable of automatically segmenting the lines and words from the document. It envisages to form the basis for building the word models and facilitate in retrieving the document that contains the query word.

4. PROPOSED METHODOLOGY

The fundamental issue in analyzing and retrieving handwritten documents lies in handling the variations in the style of writing. Among the two distinct variations of inter and intra person, the former appears to be very common as every individual has a unique writing style. It is such uniqueness in handwriting that is inherently present in every individual that enables handwriting to qualify as a biometric suitable for some person identification tasks. The benefits of indexing handwritten documents are obvious due to the ease with which a search can be made on them. However the primary task of segmentation of handwritten documents into lines and words, recognition and indexing of segmented words, quick and efficient retrieval of matching documents containing search words in an automated focus is yet to see its reality. There are a host of issues such as non-uniform spacing between words and lines in documents, inter-person and intra-person variation in writing styles, degradation in quality of document images, complexity in the strokes used in writing, and appearance of special characters in the document, that require to be addressed in the construction of an interactive domain.

The approach to word recognition in handwritten documents includes the creation of an index of words similar to the one found at the back of textbooks. It encompasses a host of activities as detailed in Figure 1. While scanning attempts to obtain a digital image of the handwritten documents, preprocessing serves to remove noise from them. The scanned output is a grey scale image with the intensity values of the pixels ranging from 0 to 255. The median filter is used to replace the intensity value of the pixels in a chosen neighbourhood with the median of the intensity values in that neighbourhood [7]. The filtered document image is converted to a binary image based on a threshold value with the intensity of each pixel value being 0 or 1, the process being referred to as binarization. The resultant binarized image contains 0s representing inked pixels and 1s representing white spaces.

The next stage is to segment the documents into lines and then lines into words, a nontrivial and crucial task when it comes to the case of handwritten documents as the writing style is not consistent throughout the document. This is followed by extracting the features that represent each word and conglomerate to create a model for each unique word in the document using all the words in each cluster. The number of models to be constructed depends on the number of unique words present in the handwritten document.

4.1. Segmentation

An accurate and precise segmentation algorithm results in the creation of models that best represent the word images. The algorithm for line segmentation based on the intensity values of the pixels is presented below:

Algorithm: Line Segmentation

Input: Filtered and binarized digital image

Output: Segmented lines

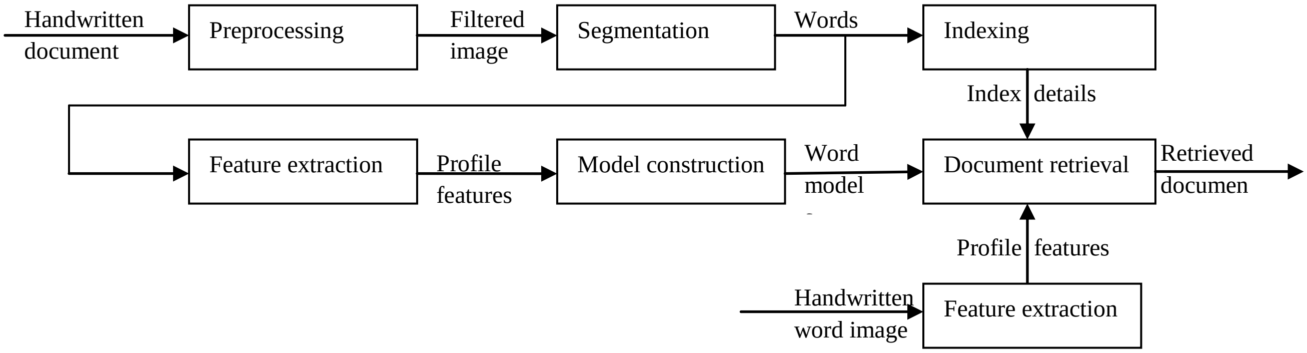


Figure 1. Components of a handwritten document retrieval system

Steps:

1. Compute the width and height of the preprocessed digital image.
2. For each row in the image, mark the row as `ink row' or `space row' by counting the number of 0s and 1s. If a row contains white spaces amounting to more than 99.5% of the entire width of the row, then mark the row as `space row', else mark it as `ink row'.
3. To mark the top and bottom boundaries of each row, identify the nature of the transition as one of the following:

(i) For a transition from space to ink row, mark the row corresponding to space as top boundary.

(ii) For a transition from space to space row, do nothing.

(iii) For a transition from ink to space row, mark the row corresponding to space as bottom boundary.

(iv) For a transition from ink to ink row, do nothing.

4. The top and bottom boundaries enclose a temporarily segmented row, temp-seg-row, and readjust all such rows to accurately enclose the text in a row, by carrying out the steps given below:

(i) For every pair of segmented rows, calculate the width of the space rows between them, to produce *sw*$_1$, *sw*$_2$,..., *sw*$_n$.

(ii) Calculate the mean space width, *msw*, as

$$msw = \frac{sw_1 + sw_2 + ...+ sw_n}{n} \qquad (1)$$

(iii) Case 1: If the width of a space row is greater than *msw*, then it is retained as a space row.

(iv) Case 2: If the width of a space row is lesser than *msw*, then merge the space row with the adjacent temp-seg-row, to form a new temp-seg-row. Repeat Step 4(iii) and 4(iv) till Case 1 becomes true and the resultant row is the desired segmented text row.

(v) Write the segmented text line as a separate image file.

The top and bottom boundary information obtained in step 3 may be readjusted so as to ignore false space rows, to correctly enclose the text in a row and follow it up with step 4. The readjustment of rows is based on a generic pattern observed in most handwritten documents. An alternative procedure is based on a threshold, predefined at the start of the segmentation process and serves to determine whether a temporarily segmented row correctly encloses text or not. However this is a static approach and may not result in good segmentation results as every writer has a unique writing style where the width of text line cannot be the same for all writers and documents.

The approach to word segmentation is carried out on the same lines as that of line segmentation, but through the columns of the image as the pivotal focus. However the segmented text lines contains spaces between words and detecting these spaces and distinguishing them from spaces between characters in a word is crucial. It is typical to Tamil language as the characters are distinctly written as isolated characters, with characters touching each other only due to the writing style of individuals. The proposed approach revolves around the fact that the space between characters in a word is generally less than the space between words and uses this pattern to mark the left and right boundaries for each word with greater precision.

In addition to this, the line segmentation algorithm is slightly modified to remove the space rows that may be present above and below the word, and results in an automatic skew corrected word. It is done by computing the height of the text line within the word image using which the top and bottom boundary of the word is determined by finding the first transition from a space row to a ink row and the last transition from an ink row to a space row. The procedure ensures even the dot, known as `pulli' that lies on top of some Tamil characters is not missed out.

4.2. Feature Extraction

The next phase is to extract the features from each segmented word that best represent the word. The features that are commonly used in document retrieval, word recognition, word spotting, and other related tasks on handwritten documents include profile features, moment based features, GSC features, and Fourier descriptors [12]. The profile features are primarily used to extract features from tamil handwritten word images and covers vertical projection profile, word profiles, and background-to-ink transitions [11]. These features are based on the intensity values of pixels denoted as *I(r,c)* where *r* and *c* correspond to the row and column of a pixel.

1. Vertical projection profile: A vertical projection profile (f_1) captures the distribution of ink along the columns and is computed by finding the sum of the intensity values in each image column separately. The vertical projection profile (*vpp*) is given by the equation.

$$vpp\,(I, c) = \sum_{r=1}^{h} (255 - I\,(r, c)) \qquad (2)$$

2. Word profiles: Word profiles capture part of the outlining shape of the word. It can also be termed as finding the upper and lower boundaries of the word image, and hence results in two features, namely, upper word profile (f_2) and lower word profile (f_3). The features are calculated by determining whether a pixel is an `ink' pixel or a `space' pixel, using a function *isInk(I,r,c)*, given by

$$isInk\,(I, r, c) = \begin{cases} 1, \; if\; I(r, c) = 0 \\ 0, if\; I\,(r, c) = 1 \end{cases} \qquad (3)$$

The upper and lower word profile uwp and lwp computation is done as

$$uwp\,(I, c) = \begin{cases} undefined, & if\; \forall_r (isInk(I, r, c) = 0 \\ \underset{r = 1, \dots, h}{argmin} (isInk\,(I, r, c) = 1), otherwise \end{cases} \qquad (4)$$

$$lwp\,(I, c) = \begin{cases} undefined, & if\; \forall_r (isInk(I, r, c) = 0 \\ \underset{r = 1, \dots, h}{argmax} (isInk\,(I, r, c) = 1), otherwise \end{cases} \qquad (5)$$

If a column does not contain ink pixels, *uwp(I,c)* and *lwp(I,c)* will be undefined for that column, and this can be redefined by linearly interpolating between the nearest defined values.

3. Background-to-ink transitions: The number of transitions from background to ink (f_4) in each column is used to capture part of the inner structure of the word.

The computed features f_1 to f_4 are normalized in the range [0..1].

4.3. Model Construction

Hidden Markov models (HMM) are found to be promising for a number of speech recognition tasks and its suitability for character or word modeling in handwriting recognition is being explored in recent times. If each handwritten word is represented by a sequence of feature vectors known as observations *O*, defined as

$$O = o_1\ , o_2\ \ , \ldots \ldots . , o_t \qquad (6)$$

where o_t is a feature vector observed at time *t*. The word image matching problem can be stated as computing

$$\underset{i}{argmax}\ P(w_i\ |O) \qquad (7)$$

A HMM is characterized by the following [13]:

1. N, the number of hidden states in the model. The individual states are indicated as S=S_1, S_2, ..., S_N, and the state at time *t* as q_t.

2. M, the number of distinct observation symbols per state. The observation symbols are denoted as $V = v_1, v_2, ..., v_M$.

3. The state transition probability distribution $A = a_{ij}$, where

$$a_{ij}\ = P\big[q_{t+1}\ =\ S_j\ \big|\ q_t\ =\ S_j\ \big], 1 \le i,j\ \le N \qquad (8)$$

4. The observation probability distribution in state *j*, *B=bj(k)*, where

$$b_j\,(k) = P\big[v_k\ at\ t\big|q_t\ =\ S_j\ \big], 1 \le j \le N\ , 1 \le k \le M \qquad (9)$$

5. The initial state distribution $\pi = \{\pi_{\ i}\}$ where

$$\pi_i\ = P[q_i\ \ =\ S_i], 1 \le i\ \ \le N \qquad (10)$$

A complete specification of a HMM is as follows:

$$\lambda = \{a_{i,j}\ , b_{j,k}\ , \pi\}, \textstyle\sum_j a_{i,j}\ = 1, \forall_i\ \sum_k b_{j,k}\ = 1, \forall_j \qquad (11)$$

The evaluation of the HMM orients to determine the probability of the observation sequence generated by a given model, and is accomplished using the forward algorithm. The decoding procedure targets to find the mst probable sequence of hidden states for a given a sequence of visible states V^T. The Viterbi algorithm finds at each time step *t*, the state that has the highest probability, thus resulting in the desired full path. The task to be accomplished in learning is to determine the model parameters from the training samples and the forward-backward algorithm or Baum-Welch algorithm is used for the purpose [8].

5. EXPERIMENTAL RESULTS

5.1. Training Phase

The proposed handwritten document retrieval system for Tamil language needs the creation of a corpus with sample handwritten document images for purpose of training and validation. A hundred people who can read, write, and speak Tamil fluently, belonging to both genders and in the age group of 14 to 40 are made to write four pages of text in Tamil language. The text includes one page each of Tamil Thai Vazhthu, five couplets of Thirukkural, one short story and one essay, the former two being poems, and the remaining two belong to the prose category. These documents are preprocessed and segmented into lines and words using the algorithm presented in Section 4.1. The word images obtained from segmentation are used for the modeling task as well as index creation.

A list of ten words from each of the four documents is chosen, thus resulting in forty words in the vocabulary pertaining to model construction. The words are chosen so as to form a representative set of the various shapes, ascent and descent present in the character set of Tamil language. The HMM with 5 states and 2 Gaussian mixtures per state is created for each word in the vocabulary, using the segmented word images collected from a number of documents. During training, the models are first initialised and the parameters of the HMM are estimated using the features f_1 to f_4 extracted from the word images. These models are then reestimated using Baum-Welch reestimation procedure to yield the models representing each word in the vocabulary.

5.2. Results of Segmentation

The system is implemented in Visual C++ and is tested for all the 40 words in the vocabulary list. Ten writers are made to write the forty words and the written words are scanned to create a test set of 40x10=400 digitized word images. The performance of the system is evaluated by providing a word image of the search word and retrieving the relevant handwritten documents containing the search word. The system computes the feature vectors of the search word followed by the most likelihood estimation for all possible models using which the word whose model likelihood is found to be the highest is selected. Based on the matched model and using the details present in the index, the appropriate document that contains the search word is retrieved and the search word is spotted within the document.

The implementation results of segmentation algorithm proposed in Section 4.1 are given in Figure 2. It is worthwhile to note that precise segmentation is the key to perfect model and index creation tasks. The accuracy of segmentation can be defined as the number of documents that are correctly segmented by the system. It can be stated as

$$A = \frac{C}{N}\ * 100 \qquad (12)$$

where A denotes the accuracy of segmentation, C the number of documents correctly segmented, and N the total number of documents subjected to segmentation.

An average of the accuracy results given in Table 1 indicates an overall segmentation accuracy of around 90%. It follows that documents in the prose category are segmented more precisely than documents in the poetry category. This can be largely attributed to the near uniform spacing that exists in essays and similar documents. However the incorrect segmentation results are mainly due to the writing style that includes adjacent lines of text touching each other. In such cases, the line segmentation algorithm treats two adjacent lines as one line segment and such incorrectly segmented portions of text can be subjected to segmentation for a second time, but by choosing a suitable threshold value for discriminating ink rows and space rows.

Table 1. Segmentation Results

#	Document	Lines per document	Words per document	Segmentation Accuracy
1	Thirukkural	10	35	86
2	Tamil Thai Vazhthu	13	37	84
3	Short story	17	80	90
4	Essay	20	87	92

5.3. Recognition and Retrieval Results

The retrieval system can be measured using the retrieval accuracy defined as

$$R = \frac{N_r}{N_t} * 100 \qquad (13)$$

where R denotes the retrieval accuracy, N_r denotes the number of search word images for which relevant documents were retrieved and N_t denotes the total number of search word images. The forty search words written by each writer forms a separate test set, the results of which are displayed in Figure 4. The retrieval accuracy computed as an average of the results obtained for the ten test sets is 80.75%. In addition to retrieving the relevant documents, the system also spots the search word in the document by making use of the details stored in the index. The snapshots of the user interface demonstrating the successful search results of the implementation of the handwritten document retrieval system are shown in Figure 3.

6. CONCLUSION

The handwritten document retrieval system for retrieving handwritten documents in Tamil language based on search words given in the form of word images has been developed. The system has been constituted to include algorithms for segmentation of handwritten text into lines and words, create HMMs to represent the words, recognize the search word and finally spot the word in the relevant document. The segmentation strategies have been based on the intensity values of the pixels in the word images. The task of document retrieval has been accomplished by word recognition using HMMs and then word spotting during document retrieval using segmentation and indexing. The performance of the handwritten document retrieval system has been computed to be 80.75% and will go a long way in enhancing the scope of the handwritten document retrieval systems.

7. REFERENCES

[1] Konstantinos Zagoris, Kavallieratou Ergina, Nikos Papamarkos. 2010. A document image retrieval system. Engineering Applications of Artificial Intelligence. Vol. 23, No.6, 872-879 .

[2] Huaigu Cao, Anurag Bhardwaj, Venu Govindaraju. 2009. A probabilistic method for keyword retrieval in handwritten document images. Pattern Recognition. Vol. 42, No.12, 3374-3382.

[3] G. Louloudis, B. Gatos, I. Pratikakis, C. Halatsis. 2009. Text line and word segmentation of handwritten documents. Pattern Recognition. Vol. 42, No.12, 3169-3183.

[4] Kaustubh Bhattacharyya, Kandarpa Kumar Sarma. 2009. ANN-based innovative segmentation method for handwritten test in Assamese. Intl. Journal of Computer Science Issues. Vol. 5, 9-16.

[5] Million Meshesha, C. V. Jawahar. 2008. Matching word images for content based retrieval from printed document images. International Journal on Document Analysis and Recognition. Vol. 11, No.1, 29-38.

[6] Jose A. Rodriguez, Florent Perronnin, Gemma Sanchez, Josep Llados. 2008. Unsupervised writer style adaptation for hand written word spotting. 19th International Conference on Pattern Recognition. 1-4.

[7] Rafael C. Gonzalez, Richard E. Woods. 2008. Digital Image Processing. Prentice Hall.

[8] Richard O. Duda, Peter E. Hart, David G. Stork. 2007. Pattern Classification. Wiley-India.

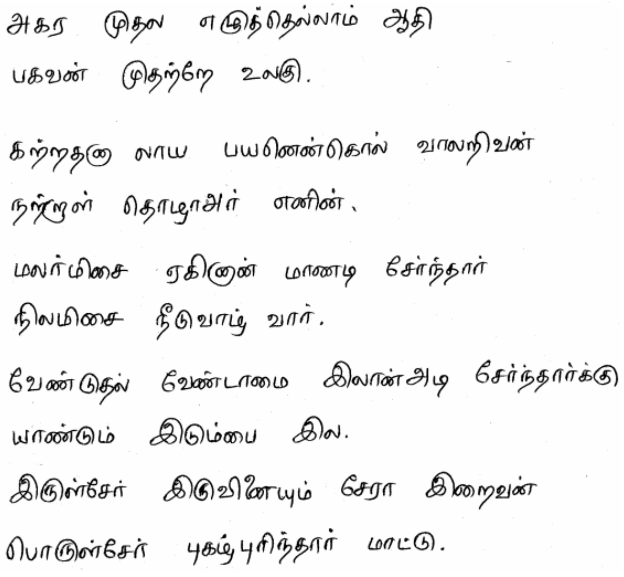
[9] Gregory R. Ball, Sargur N. Srihari, Harish Srinivasan. 2006. Segmentation-based and Segmentation-free methods for spotting handwritten Arabic words. 10th International Workshop on Frontiers in Handwriting Recognition. 53-58.

[10] Sargur Srihari, Chen Huang, Harish Srinivasan. 2005. A search engine for handwritten documents. In Proc. of Document Recognition and Retrieval. 66-75.

[11] Toni M. Rath, R. Manmatha. 2003. Features for word spotting in historical manuscripts. ICDAR. 218-222.

[12] O Due Trier, Anil K. Jain, Torfin Taxt. 1996. Feature extraction methods for character recognition: A Survey. Pattern Recognition. Vol. 29, No.4, 641-662.

[13] Lawrence R. Rabiner. 1989. A tutorial on hidden Markov models and selected applications in speech recognition. Proceedings of the IEEE. Vol. 77, No.2, 257-285.



அகர முதல எழுத்தெல்லாம் ஆதி
பகவன் முதற்றே உலகு.

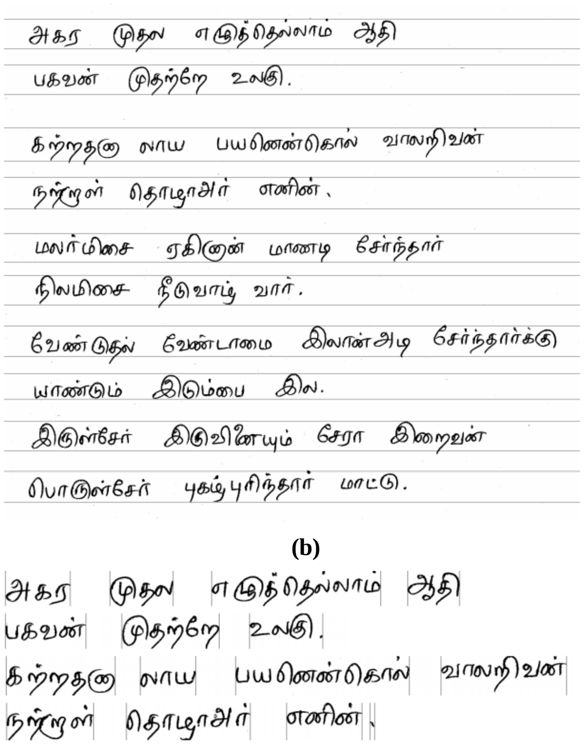
கற்றதனா லாய பயனென்கொல் வாலறிவன்
நற்றாள் தொழாஅர் எனின்.

மலர்மிசை ஏகினான் மாணடி சேர்ந்தார்
நிலமிசை நீடுவாழ் வார்.

வேண்டுதல் வேண்டாமை இலான்அடி சேர்ந்தார்க்கு
யாண்டும் இடும்பை இல.

இருள்சேர் இருவினையும் சேரா இறைவன்
பொருள்சேர் புகழ்புரிந்தார் மாட்டு.

(a)



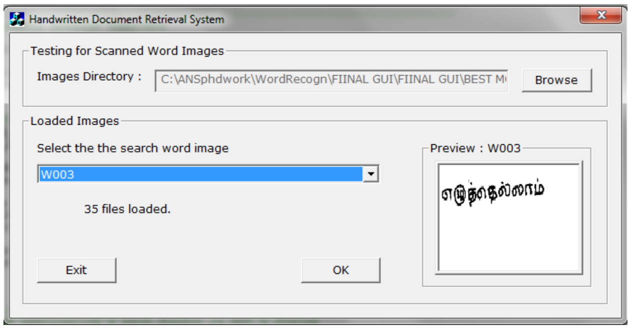
அகர முதல எழுத்தெல்லாம் ஆதி
பகவன் முதற்றே உலகு.
கற்றதனா லாய பயனென்கொல் வாலறிவன்
நற்றாள் தொழாஅர் எனின்.
மலர்மிசை ஏகினான் மாணடி சேர்ந்தார்
நிலமிசை நீடுவாழ் வார்.
வேண்டுதல் வேண்டாமை இலானடி சேர்ந்தார்க்கு
யாண்டும் இடும்பை இல.
இருள்சேர் இருவினையும் சேரா இறைவன்
பொருள்சேர் புகழ்புரிந்தார் மாட்டு.

(b)

அகர முதல எழுத்தெல்லாம் ஆதி
பகவன் முதற்றே உலகு.
கற்றதனா லாய பயனென்கொல் வாலறிவன்
நற்றாள் தொழாஅர் எனின்.

(c)

Figure 2. Segmentation of handwritten Tamil document. (a) A sample handwritten Tamil document (b) Output of line segmentation (c) A portion of the output of word segmentation



Handwritten Document Retrieval System

Testing for Scanned Word Images

Images Directory : C:\ANSphdwork\WordRecogn\FIINAL GUI\FIINAL GUI\BEST M

Loaded Images

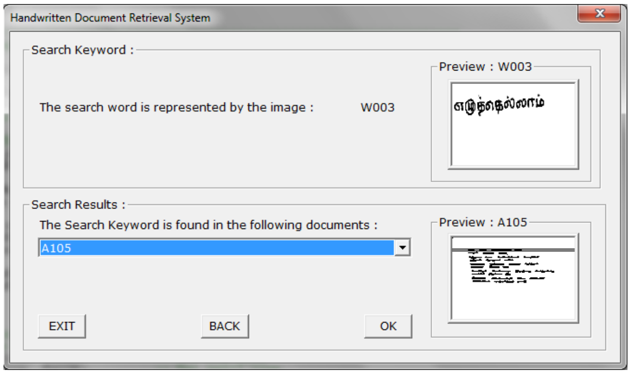
Select the the search word image: W003

35 files loaded.

Preview : W003

எழுத்தெல்லாம்

(a)



Handwritten Document Retrieval System

Search Keyword :

The search word is represented by the image : W003

Preview : W003

எழுத்தெல்லாம்

Search Results :

The Search Keyword is found in the following documents : A105

Preview : A105

(b)

Figure 3. Snapshot of the user interface of the handwritten document retrieval system. (a) Selection of search word image (b) Spotting the search word in the retrieved document

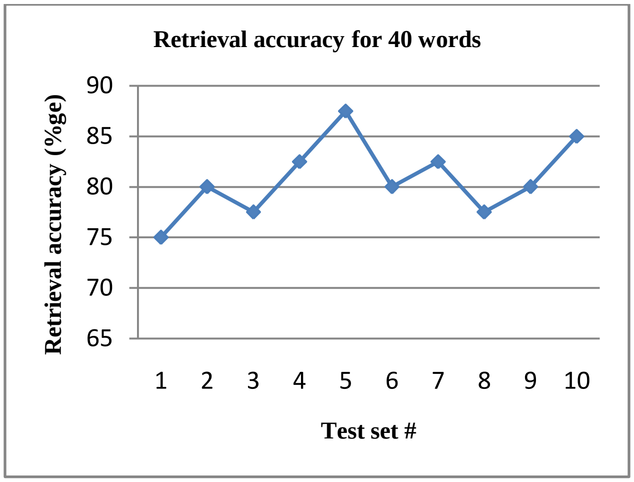


Figure 4. Retrieval accuracy obtained for 40 words from 10 test sets