International Journal of Applied Information Systems (IJAIS) ISSN : 2249-0868
Foundation of Computer Science FCS, New York, USA
Volume 1 - No. 1, November 2012 - www.ijais.org

Query Word Image based Retrieval Scheme for Handwritten Tamil Documents

A N. Sigappi
Department of Computer Science and Engineering
Annamalai University
Annamalainagar-608002, India

S. Palanivel
Department of Computer Science and Engineering
Annamalai University
Annamalainagar-608002, India

ABSTRACT

This paper brings out an autoassociative neural network (AANN) based information retrieval mechanism to locate handwritten documents from a literary collection in Tamil language corresponding to query word images. The strategy extends to create models for the chosen search word images, evolve a methodology to identify the search word and subsequently retrieve the relevant documents. AANN emphasises a training procedure through an appropriate combination of units in the layers of the network to arrive at a suitable model for each word in the vocabulary. The training phase orients to segment the digitized text documents into lines and words, extract profile and moment based features from the words and articulate an index of words. The features computed based on the intensity values of the pixels cater to accrue the nuances of the strokes in the characters. The experimental results obtained for an index of words elaborate the astuteness of the scheme and its retrieval accuracy.

General Terms:

Handwritten documents, information retrieval

Keywords:

Segmentation, profile features, moment based features, autoassociative neural networks

1. INTRODUCTION

The storage of documents in digital formats evinces a greater interest in recent times due to the availability of enormous data storage capabilities at significantly reduced costs. It attempts to revolutionize the appearance and feel of libraries across the world in improving accessibility to information. The libraries are in possession of traditional paper based documents, modern born-digital documents, and digitized documents. The born-digital documents are created in digital formats, and the tools and techniques for indexing, searching and retrieving such documents have matured. The digitized documents contain information that has been converted from a physical medium, like paper, into a digital form. Searching and retrieving such digitized documents, often available in the form of scanned digital images of paper based machine printed and handwritten documents, still remains a daunting task for researchers. Though retrieval challenges for languages like English are adequately addressed by researchers, comprehensive strategies that ensure reliable handwritten document retrieval results for regional languages like Tamil are still in their infancy.

Tamil is one of the longest surviving classical languages in the world, whose orthography consists of twelve vowels, eighteen consonants, a special letter called 'ayutham', and two hundred and sixteen characters derived by a combination of the vowels and consonants. It is a syllabic language with a total of two hundred and forty seven characters that are typically written in isolation and the direction of writing is from left to right in horizontal lines. The number of strokes, the directions of strokes and the order of strokes for each character vary widely from person to person. The inherently complex characteristic features of Tamil language along with the nonavailability of an index of keywords, and the difficulties in digitizing the aged documents necessitate the development of algorithms to address them.

2. RELATED WORK

An adaptive run length smoothing algorithm has been used to handle the problem of complex and dense document layout in the segmentation of document pages resulting from the digitization of historical machine printed sources [7]. A technique for keyword guided word spotting in historical printed documents has been suggested to initialize the word retrieval mechanism through the creation of synthetic word data combined with hybrid feature extraction [5]. A system for recognizing handwritten manuscripts based on hidden Markov models supported by a language model has been presented. This system has been found to investigate two approaches, one that involves training the recognizer from scratch, and the other that adapts it from a recognizer trained on a general offline handwriting database [3]. A document image retrieval system that performs word matching directly in the document images has been developed and evaluated on a collection of documents created from various texts with noise added in them [14]. The experimental results for Tamil online handwritten character recognition using HMM and statistical dynamic time warping (SDTW) as classifiers have been reported [4].

An analytical handwritten word recognition system combining neural networks and hidden Markov models (HMM) has been described on isolated french word images. The approach presented has been found to use the neural network outputs as observation probabilities and train the neural network to reject non-characters implicitly, without explicitly creating junk samples [11]. An autoassociative neural network has been designed to model both the classification and the retrieval problem, using a Hebbian association rule and a hyperbolic tangent activation function as the learning process [1]. The task of text document retrieval has been attempted using linear and nonlinear hebbian neural networks and linear autoassociative neural network and found to reduce the dimension of the document search space while preserving the retrieval accuracy [10].

3. PROBLEM DEFINITION

There is an inert need to preserve and conserve valuable collections of handwritten documents by way of generation of integrated and effective approaches for storing, searching, and retrieving them. The pattern classification techniques offer a significant role in creating appropriate models with a view to identify the words in Tamil language. The key perspective orients to developing an information retrieval mechanism endowed with an ability to recognize query word images and extract handwritten

documents that assuages to evaluate its performance through a sample set of documents.

4. PROPOSED METHODOLOGY

The proposed scheme consists of the following steps: digitization, preprocessing, segmentation, index creation, feature extraction, model construction, word recognition, and document retrieval. The digitization process refers to scanning and converting the handwritten documents into digital documents in image formats. It is followed by the preprocessing stage which includes noise removal and image binarization. Noise removal attempts to clean up the document image using a median filter that causes relatively less blurring of edges while removing the noise that may be present due to writing characteristics or scanning [2]. The image binarization converts the denoised image into a binary form, with 0's representing the foreground inked pixels and 1's representing the background white pixels using a suitable threshold value. The segmentation phase involves segementing the document into lines and the lines into words, and each unique word is included in an index to aid in the search task. The features that characterize the word are acquired from the segmented word and using which an autoassociative neural network model is constructed to represent each unique word. It is trained to recognize the search word by matching it against the available word models and subsequently retrieve the relevant documents using the recognized word and the index of keywords.

4.1 Segmentation

The task of segmenting the document images into lines and thereafter words is critical because of the large variation in writing styles that exists within and between documents written by same and different authors. Nonuniform spacing between lines and words, skewed writing style, discontinuities in the strokes, and overlapped strokes tend to complicate the task of segmentation. The line segmentation algorithm is designed to find the top and bottom boundaries of each text row based on the number of noninked pixels in a row. The width of the space rows between the text rows across the entire document is used to compute the width of the space row. The computed width is used as a threshold parameter to merge a space row with its adjacent text row if required owing to the writing style and the inherent nature of the Tamil characters that form the word. The advantage of this algorithm lies in the automatic readjustment of text lines based on the writing style found in the documents and hence can be effective for most documents even if the width of a text line is not the same in all documents. A similar approach can be followed for segmenting the words from the text lines, but working through the columns of the image instead of rows and marking the left and right boundaries of each word.

4.2 Feature Extraction

The goal of feature extraction is to represent the object to be recognized in terms of values that are very similar for objects in the same category, but are very different for objects in different categories. It is indeed difficult to characterize the handwritten words in terms of values, owing to the complexity and similarity that exists in the strokes of the characters in the word. However, profile features, moment based features, GSC features, and Fourier descriptors are commonly used in hand written word recognition, document retrieval and other related tasks [12]. The profile features [9] include vertical projection profile, upper and lower word profiles, and background-to-ink transitions. It serves to characterise the word image using the transitional distributional way of representing the word images and in certain cases turn out to be inadequate to distinguish the word images. Therefore if a statistical mechanism involving standard deviation, skewness, and kurtosis [8] through which the shape of the images be extracted, it can only add substance to the recognition methodology. The profile and moment features are gathered 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.

4.3 Model Construction

Autoassociative neural network models (AANNs) are feed forward neural networks that can perform an identity mapping of the input space [6] and used in several image and speech processing applications [13]. The architecture of a five layer AANN model shown in Fig. 1 comprises of an input layer, three hidden layers, and an output layer. In this network, the number of units in the input and output layers are identical. The second and fourth layers are referred to as the mapping and demapping layer respectively and their processing units are nonlinear. The third layer is called as the bottleneck or compression layer in which the units can be either linear or nonlinear. The number of units in the mapping and demapping layers is greater than the input layer whereas the compression layer has fewer units than the input layer.

The AANN is trained by optimizing the weights so that the reconstructed outputs match the input and the sum of the squared errors is minimized upon completion of training. The nonlinear output function used for each unit is $tanh(s)$, where s is the activation value of the unit. The network is trained using the backpropogation algorithm and the error for each input data point i is plotted in the form of a probability surface as $p_i = exp(\frac{-e_i}{\alpha})$ where α is a constant. The shape of the error surface is attributed to the constraints imposed by the network and facilitates the AANN to capture the distribution of the input data.

5. EXPERIMENTAL RESULTS

5.1 Network Training

A corpus is created with four hundred sample handwritten document images collected from a hundred people who can read, write, and speak Tamil language fluently, belonging to both genders and in the age group of fourteen to forty. 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 given in 4.1 and the resultant word images are used for the purpose of model construction and index creation.

A select list of forty words picked from the four documents forms the vocabulary for model construction. The words are chosen so as to form a representative set of the various strokes present in the character set of Tamil language. The segmented word images collected from a number of documents are used to create an AANN model for each word in the vocabulary. The structure of the AANN model used in the experiments is chosen by systematically varying the number of units in the second and third layers and by varying the number of epochs required for training the network. The entire network is trained by presenting the seven dimensional feature vector as input as well as desired output and the backpropogation algorithm is used to adjust the weights of the network so as to minimize the mean square error for each feature vector.

5.2 Testing

The document retrieval methodology is validated for all the forty words in the vocabulary list by creating ten test sets with forty digitized word images in each set. It is executed by providing a word image of the search word and retrieving the relevant handwritten documents containing the search word. The feature vector extracted from the search word image (s) is presented to each

of the AANN models and the output (o) obtained from each model is compared with the input to compute the normalized squared error *(e)*. The normalized squared error obtained for (s) is given by

$$e = \frac{\|s - o\|^2}{\|s\|^2} \quad (1)$$

This normalized squared error (e) is translated to a confidence score (C) as $C = exp(-e)$ and the confidence score is thus calculated for each model and the model that yields the highest confidence score is selected. The result coupled with the details present in the index aids in the retrieval of the appropriate document that contains the search word and thereafter spotted precisely within the retrieved document.

The performance of the retrieval scheme is measured using the retrieval accuracy defined as the ratio of the number of search word images for which relevant documents were retrieved to the total number of search word images. The retrieval accuracy is calculated for the test sets contributed by ten different writers and using which the overall retrieval accuracy is computed as an average of the results obtained for the ten test sets.

The structure of the network and the choice of features influence the performance of the AANN based retrieval scheme. The number of units in the input and output layers are required to be varied in order to strike at the best possible combination to accrue enviable results. The results obtained using profile and moment based features over a viable range of units in the third layer and in the second layer of the network are depicted in Fig. 2 and Fig. 3 respectively and endure to conclude an appropriate choice for a structure as $7L\ 14N\ 5N\ 14N\ 7L$.

The results obtained for the ten test sets using different sets of features is shown in Fig.4. An overall retrieval accuracy calculated from these results is given in Fig.5. A retrieval accuracy of 79% reported corresponds to a model based on combination of both profile and moment based features. However it is observed that the accuracy drops down to around 70% when the experiments are performed using only profile features, thus highlighting the need for the inclusion of a statistical orientation. Owing to the statistical nature of the identification procedure the moment based values intrude a higher signifcance and govern the cohesiveness of the entire strategy.

6. CONCLUSION

An AANN based approach for retrieving handwritten Tamil documents for a limited vocabulary of search words has been pronounced in this paper. The scheme has been formulated to aggregate algorithms for segmentation of handwritten text into lines and words and creation of AANN models to represent the words. The segmentation strategies have been designed to beehive moment based characteristics along with profile features based on the pixel intensities of word images. The task of document retrieval has been accomplished by word recognition using AANNs and then word spotting during document retrieval using segmentation and indexing. The performance of the handwritten document retrieval scheme has been found to be 79% and contemplate to address the complexities prevalent in the handwritten document retrieval domain.

7. REFERENCES

[1] Guy Desjardins, Robert Proulx, and Robert Godin. An auto-associative neural network for information retrieval. In *International Joint Conference on Neural Networks, IJCNN 2006, part of the IEEE World Congress on Computational Intelligence*, pages 3492–3498, July 2006.

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

[3] Emanuel Indermhle, Marcus Eichenberger-Liwicki, and Horst Bunke. Recognition of handwritten historical documents: HMM-adaptation vs. writer specific training. In *11th Int. Conference on Frontiers in Handwriting Recognition*, pages 186–191, 2008.

[4] Shashi Kiran, Kolli Sai Prasada, Rituraj Kunwar, and A. G. Ramakrishnan. Comparison of HMM and SDTW for tamil handwritten character recognition. In *IEEE International Conference On Signal Processing and Communications*, pages 1–4, 2010.

[5] T. Konidaris, B. Gatos, K. Ntzios, I. Pratikakis, S. Theodoridis, and S. J. Perantonis. Keyword-guided word spotting in historical printed doucments using synthetic data and user feedback. *Intl. Journal on Document Analysis and Recognition*, 9(2):167–177, April 2007.

[6] Mark A. Kramer. Nonlinear principal component analysis using autoassociative neural networks. *AIChE Journal*, 37(2):233–243, February 1991.

[7] Nikos Nikolaou, Michael Makridis, Basilis Gatos, Nikolaos Stamatopoulos, and Nikos Papamarkos. Segmentation of historical machine-printed documents using adaptive run length smoothing and skeleton segmentation paths. *Image and Vision Computing*, 28(4):590–604, 2010.

[8] William H. Press, Saul A. Teukolsky, William T. Vetterling, and Brian P. Flannery. *Numerical recipes in C*. Cambridge University Press, 2002.

[9] Toni M. Rath and R. Manmatha. Features for word spotting in historical manuscripts. In *ICDAR*, pages 218–222, 2003.

[10] Lenka Skovajsova. Text document retrieval by feed forward neural networks. *Information Sciences and Technologies Bulletin of the ACM Slovakia*, 2(2):70–78, 2010.

[11] Yong Haur Tay, Pierre Michel Lallican, Marzuki Khalid, Stefan Knerr, and Christian Viard-Gaudin. An analytical handwritten word recognition system with word-level discriminant training. In *ICDAR*, pages 726–730, 2001.

[12] O Due Trier, Anil K. Jain, and Torfin Taxt. Feature extraction methods for character recognition: A survey. *Pattern Recognition*, 29(4):641–662, Feb 1996.

[13] B. Yegnanarayana and S. P. Kishore. AANN: an alternative to GMM for pattern recognition. *Neural Networks*, 15(3):459–469, April 2002.

[14] Konstantinos Zagoris, Kavallieratou Ergina, and Nikos Papamarkos. A document image retrieval system. *Engineering Applications of Artificial Intelligence*, 23(6):872–879, Sept 2010.

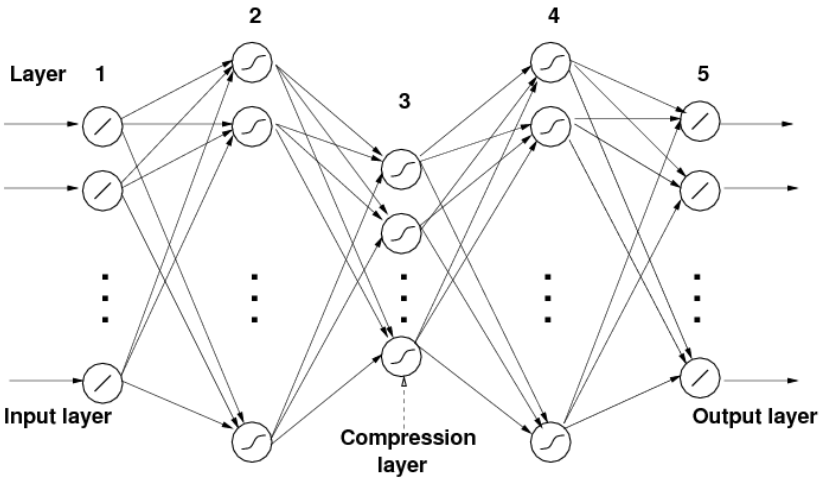

Fig. 1. A five layer Autoassociative neural network model.

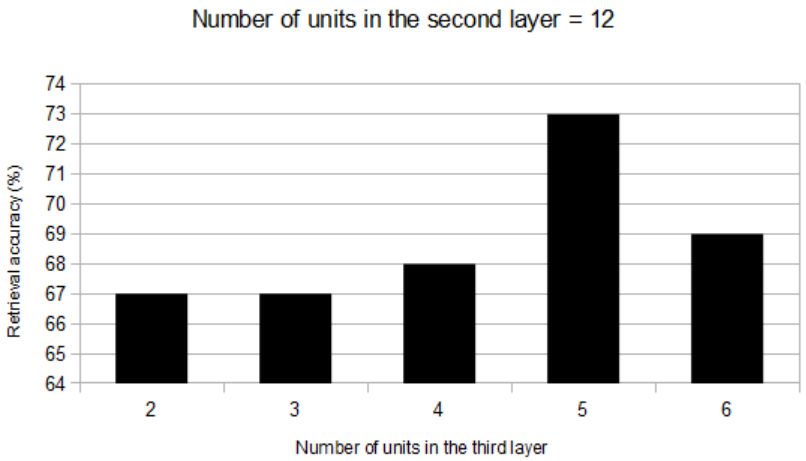

Fig. 2. Retrieval accuracy obtained by varying the units in the third layer of AANN.

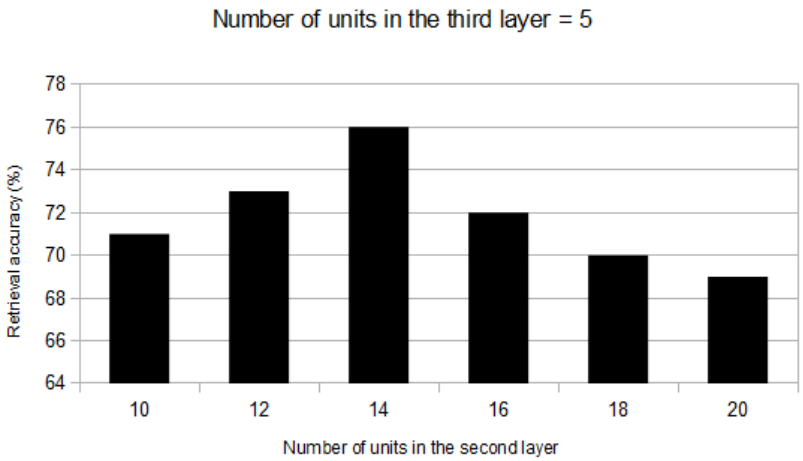

Fig. 3. Retrieval accuracy obtained by varying the units in the second layer of AANN.

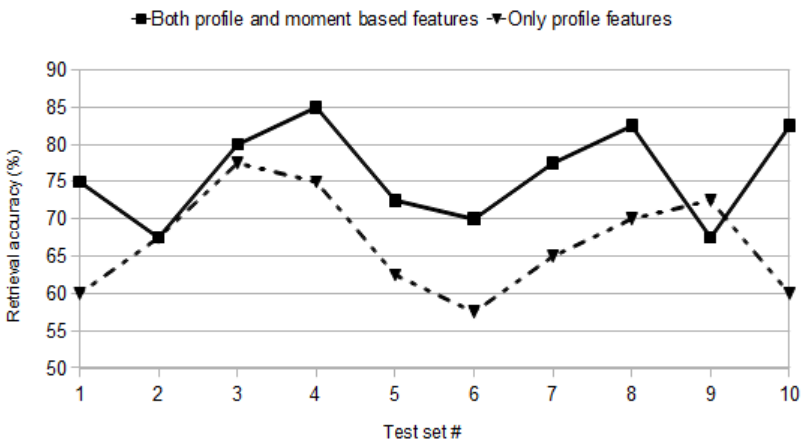

Fig. 4. Retrieval accuracy obtained for 10 test sets for various features.

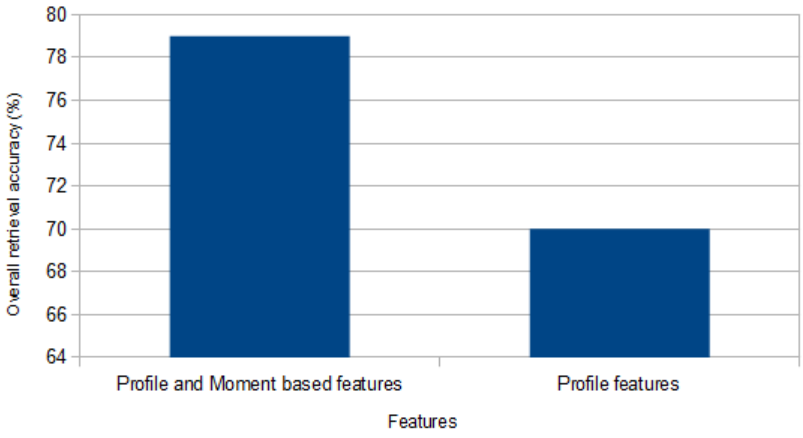

Fig. 5. Overall retrieval accuracy obtained for various features.