An Adaptive Character Recognizer for Telugu Scripts using Multiresolution Analysis and Associative Memory

ARUN K PUJARI,
AI Lab, University of Hyderabad,
Hyderabad, 500 046 INDIA
akpcs@uohyd.ernet.in

C DHANUNJAYA NAIDU,
Department of ECE, VNR-VJIET,
Hyderabad, 500 072 INDIA
cdnaidu@yahoo.com

B C JINAGA
Department of ECE,
J N Technological University,
Hyderabad, 500 072 INDIA

Abstract

The present work is an attempt to develop a commercially viable and a robust character recognizer for Telugu texts. We aim at designing a recognizer which exploits the inherent characteristics of the Telugu Script. Our proposed method uses wavelet multiresolution analysis for the purpose extracting features and associative memory model to accomplish the recognition tasks. Our system learns the style and font from the document itself and then it recognizes the remaining characters in the document. The major contribution of the present study can be outlined as follows. It is a robust OCR system for Telugu printed text. It avoids feature extraction process and it exploits the inherent characteristics of the Telugu character by a clever selection of Wavelet Basis function which extracts the invariant features of the characters. It has a Hopfield-based Dynamic Neural Network for the purpose of learning and recognition. This is important because it overcomes the inherent difficulties of memory limitation and spurious states in the Hopfield Network. The DNN has been demonstrated to be efficient for associative memory recall. However, though it is normally not suitable for image processing application, the multi-resolution analysis reduces the sizes of the images to make the DNN applicable to the present domain. Our experimental results show extremely promising results.

1. Introduction

Optical character recognition (OCR) has been one of the most well studied problems in Pattern Recognition. Today, reasonably efficient and inexpensive OCR packages are commercially available to recognize printed texts in widely used languages such as English, Chinese, and Japanese etc. A varieties of techniques of Pattern Recognition such as Template Matching, Neural Networks, Syntactical Analyses, Wavelet Theory, Hidden Markov Models, Bayesian theory etc have been explored to develop robust OCRs for different languages such as Latin, Chinese (Kanji) [Wang, 2001], [Casey, 66], Hangul (Korean) scripts [Kim, 2001], Arabic script [Cheung, 2001] also. There have also been some attempts to develop OCRs for some Indian languages like Devnagari[Sinha, 79], Bengali[Chaudhuri, 97], Telugu [Rajasekhar, 77], [Sukhaswamy, 95] and Gujarati. Notwithstanding the importance and the need, this problem is not adequately investigated by the researchers but for some isolated investigations. From the research investigation point of view the following questions can be asked.

- Can any of the existing OCR systems be used in its present form (or with modification) for recognition of any of Indian language texts? This is apparently not feasible because the construction of words follows a complex rule and not simply the concatenation of characters. Readers are referred to [Sukhaswamy, 95] and [Negi, 01] for detail account of the characteristics of Telugu language.
- Can there be some specific structural features that can be identified for any of the Indian scripts? Some work in this area have been done for Devnagari Script [Sinha, 79], [Sinha, 87], for Bangla Script [Chaudhuri, 98], for Telugu script [Sukhaswamy, 95][Negi, 01]. Most of the attempts of OCRs for Indian languages fall into this category.

The present work is an attempt to develop a commercially viable and a robust character recognizer for Telugu texts. It may be noted here that Telugu is the second widely used Indian language and has a script that is different from the Devnagari script used for Hindi language or Bangla script that is used for Bengali language.

The proposed method uses wavelet multi-resolution analysis for the purpose of extracting features and associative memory model to accomplish the recognition tasks. An important point to note here that we exploit the inherent features of Telugu scripts for recognition. The major contribution of the present study can be outlined as follows.

- It avoids feature extraction as it does not explicitly identify the features that are to be extracted. It totally relies on Wavelet multi-resolution analysis for capturing the distinctive characteristics.
- It exploits the inherent characteristics of the Telugu character. This is done by a clever selection of Wavelet Basis function which extracts the invariant features of the characters. Our experimental results show extremely promising results. The experiments also indicate that the proposed technique does not give good results for English text. This establishes that the proposed system is able to encode the specific nature of the Telugu script for the purpose of recognition
- It has a Hopfield-based Dynamic Neural Network for the purpose of learning and recognition. This is important because it overcomes the inherent difficulties of memory limitation and spurious states of the conventional Hopfield Network. The DNN has been demonstrated to be efficient for associative memory recall.

2 Earlier Works

The readers are referred to survey papers and text books for good exposure of the OCRs [Trier, 96], [Mori, 92], [Elliman, 90], [Dunn, 92], [Lam, 92], [Fu, 82], [Duda, 73], [Fukunga, 90], [Devijver, 82], [Ullmam, 73], [Govindan, 90], [Nagy, 92], [Suen, 80], [Mantas, 86], [Impedevo, 91], [Mohiuddin, 94], [Wilson, 93]. The different techniques of OCR can be broadly classified into two methods- Feature Mapped Recognition and Image Mapped Recognition. In the Feature Mapped Recognition, the recognition task is accomplished by extracting certain primitives or distinctive features. The individual characters are recognized based on a decision function that decides the presence and absence of different primitive components in the character. In the Image Mapped approach the identification and the extraction of features are implicit processes within the recognition process. Neural Network based OCRs mostly follow this principle [Burr, 88], [Fukushima, 91], [Perantonis, 92], [Guyon, 91], [Schwenk, 95], [Kunihito, 92], [Hubrig-Schaumburg, 92], [Sabourin, 92], [Grother, 92], [Lovell, 94],

The first reported work on OCR of Telugu Character is by Rajasekharan et al [Rajasekhar, 77]. This work identifies 50 primitive features and proposes a two-stage syntax-aided character recognition system. Primitives are joined and superimposed appropriately to define individual characters. Firstly, these primitives are recognized by a *Sequential Template Matching* mechanism. The basic letters are recognized by a process called *On The Curve Coding*. Though the experimental results are sufficiently promising, its validity in a practical OCR is not investigated. In [Sukhaswamy, 95] a neural network based system was proposed to recognize Telugu script. An extensive study is undertaken to identify the structural chacteristics of Telugu script and the distinct symbols of the Telugu language are categorized based on their relative size. The authors propose neural network architecture and investigate different learning techniques to explore the recognition capability of such the network. It is an Image Mapped system and it is demonstrated that the proposed network can yield extremely efficient recognition. The work is a demonstration of robustness of a hierarchical Hopfield Network for the purpose of recognition of noisy Telugu Characters. The present work is inspired by the work reported in [Sukhaswamy, 95]. Recently in [Negi, 2001], a simple but robust Telugu OCR is proposed based on template matching of the distinct connected components using fringe distance metric. Very encouraging experimental results on certain fonts are reported. The techniques can perhaps be applicable to other Indian Languages also but it is not clear whether the specific features of Telugu scripts are exploited.

3 The Proposed Method

In the present work we propose a new method of recognizing Telugu characters. We make use of wavelet analysis to capture the invariant features of the Telugu scripts. The basic characters of this script are inherently circular in nature. Unlike Latin script, Chinese script, Devnagari or Bangla script, Telugu characters very rarely contain horizontal, vertical or diagonal line. We observe that the Telugu characters are obtained by joining circular shapes (full or partial) of different sizes with some modifiers. The modifiers are either of circular shape or oblique linear strokes. This observation is the primary motivation of the present study. The Wavelet representation encodes the *average image* and the *detail images* as we transform the image for coarser resolutions. The *detail images* encode the directional features of vertical, horizontal and diagonal direction whereas the *average image* retains the average features. Thus the average image of a Telugu symbol can retain its salient structure even at coarser resolution. However not all types of basis functions are able to preserve these features in the present context. We tried a set of scaling functions and noticed that their behaviour differ. We report this observation in a later section. Another advantage of using wavelet representation is that the preprocessing stages of contouring, thinning or edge detection are no longer necessary. We apply wavelet transform to the grey scale image and use the process of binarization on wavelet coefficient of the average image. We describe below the stages of the proposed process.

3.1 Scanning and Size Normalization

We first scan a text as grey-scale image of appropriate resolution (at present we scan with 300 dpi resolution). The text may consist of running text material of multiple pages. In this present study we assume that the text is free of mathematical symbols, figures, or tables. The digitised image is segmented to extract individual characters. We follow the projection (horizontal, for line segmentation and vertical for the word segmentation) techniques proposed in [Sukhaswamy, 95]. Other techniques such as connected component approach [Negi, 2001] can also be adopted. Each individual characters can be of different size and hence, we convert them to images of fixed size by zero-padding techniques. We observe that though sizes vary for different characters, a frame of 32x32 is appropriate as the normalized size to contain all distinct symbols. The proposed method would work for any other size.

3.2 Wavelet Analyses

Wavelet representation gives information about the variations in the image at different scales. A high wavelet coefficient at a coarse resolution corresponds to a region with high global variations. The idea is to find relevant point to represent this global variation by looking at wavelet coefficients at finer resolutions. The (forward) Wavelet transforms can be viewed as a form of subband coding with a lowpass filter (*H)* and a highpass filter *(G)* which split a signal's bandwidth in half.

The one-dimensional forward wavelet transform of a signal s is performed by convolving s with both *H* and *G* and down sampling by 2. Figure 1 illustrates a single two-dimensional forward wavelet transform of an image, which is accomplished by two separate one-dimensional transforms. The image *f(x,y)* is first filtered along the *x*-dimension, resulting in a lowpass image $f_L(x, y)$ and a highpass image $f_H(x, y)$. The sampling is accomplished by dropping every other filtered value. Both f_L and f_H are then filtered along the *y* dimension, resulting in four subimages: f_{LL}, f_{LH}, f_{HL} and f_{HH}. Once again, we can downsample the subimages by 2, this time along the *y*-dimension. As illustrated in Figure 1, the 2-D filtering decomposes an image into an *average signal* f_{LL} and three *detail signals* which are directionally sensitive: f_{LH} emphasizes the horizontal image features, f_{HL} the vertical features, and f_{HH} the diagonal features. This process reduces a 32x32 image to a set of four 8x8 images of which we consider only one, namely the *average image*.

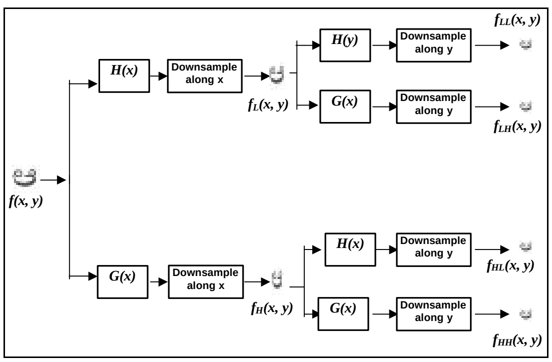

Figure 1 Signature computation with wavelet

3.3 Signature Computation

After three levels of decomposition we take the average image of size 8x8. We convert the images to binary images by a process of thresholding where the mean value of the grey level is selected as the threshold. The resulting bit string of 64 bits is termed as the signature of the input symbol. At this stage we can frame the character recognition problem as follows. We are given with a training set of 64-bit signatures representing the distinct character set, and we are to recongnise each of the input of 64-bit string This represents an unknown character of the page. Thus we have a typical associative memory recall problem. However, the well known Hopfield model suffers from memory limitation and spurious state and hence, is not suitable in the present context. This is particularly so due to large number of distinct symbols. Though Telugu also has 52 basic characters the number distinct symbols are very large. Recently, in [Rao, 99] a new neural network model is proposed which has many novel features such as order-sensitive learning, pruning and spurious-free recall of very large memory.

4. Dynamic Neural Network (DNN)

In [Rao, 99] a new architecture of neural network is proposed as Dynamical Neural Network. This is a composite structure wherein each node of the network is a Hopfield network by itself. The Hopfield network employs a new learning rule and hence converges to the user-specified stable states without admitting any spurious states. The network is called to be dynamic as it changes its architecture during the process. The DNN contains a set of basic nodes which are identical Hopfield networks. However, during the learning phase, they acquire different synaptic weights. The basic nodes are grouped together in a hierarchical organization. Each group has a designated basic node called the *leader*. When a test pattern is presented to DNN, it is presented to all the basic nodes at the lowest level of the hierarchy of nodes. Each node reaches its own stable state based on the common input and individual synaptic weights. These nodes transmit their stable states to their respective leader. At this stage the DNN adopts a pruning mechanism and retains only the leader nodes. These leader nodes are treated as the basic node of the next level of the hierarchy and they in turn send the resulting states to the leader nodes at the next level of the hierarchy after reaching the stable state. The process proceeds in this way till the whole network reaches a single stable state. In one cycle, the available basic nodes carry out the state-transition function with the given synaptic matrices and in the next cycle these nodes communicate among themselves to change the synaptic weights. At this stage the network is pruned to retain only the leader nodes of the current level of the hierarchy. In [Rao, 99] it is shown that by this process the network can accomplish very efficient associative recall without any spurious states or any sort of memory limitations, which are two main drawbacks of the Hopfield model of the neural networks. It is however not suitable for the cases where the database size is small but each individual tuple in the database is large. This is typically the case with image processing applications where each individual image is large in size and the image database need not contain very large number of distinct images. In our present study, we get over this difficulty by using the wavelet transform of the images to reduce the size of distinctive features of the images. However, we improve the DNN model by retaining back the pruned nodes and reusing them in subsequent iterations.

5 Experimental Analysis

We experimented with several pages scanned from multiple sources. A sample page of the scanned image is shown in Figure 2. We performed several experiments on different fonts from various books (we have selected mostly children's' books). The documents were scanned on an HP scanjet 3300C scanner at different resolutions. For the purpose of signature computation, we tried a variety of scaling functions such as Haar basis, Daubechie's basis, the ones proposed by Battle-Lemarie [Mallat, 89], and by Adelson and Simoncelli [Pentland, 92]. We observe that Battle-Lemarie filter is able capture the characteristics of Telugu features very efficiently. If the signatures are computed using Haar and other well known features the recognition accuracy is very low. However, the Adelson-Simoncelli filter gives a little better result. Thus we report results of only these two filters. The set of filters that are used in our experiments are given below.

Battle-Lemarie Filter [Mallat, 89]

h(n): 0.30683, 0.54173, 0.30683, -0.035498, -0.077807, 0.022684, 0.0297468, -0.0121455, -0.0127154, 0.00614143, 0.0055799, -0.00307863, -0.00274529, 0.00154264, 0.00133087, -0.000780461, 0.000655628, 0.0003955934

g(n): 0.541736,-0.30683, -0.035498, 0.077807, 0.022684, -0.0297468, -0.0121455, 0.0127154, 0.00614143, -0.0055799, -0.00307863, 0.00274529, 0.00154264, -.00133087, -0.000780461, 0.000655628, 0.0003955934, 0.000655628

Adelson- Simoncelli Filter [Pen, 91]

h(n): 0.7973934, -0.41472545, -0.073386624, 0.060944743, 0.02807382

g(n): 0.7973934, -0.41472545, -0.073386624, 0.060944743, 0.02807382

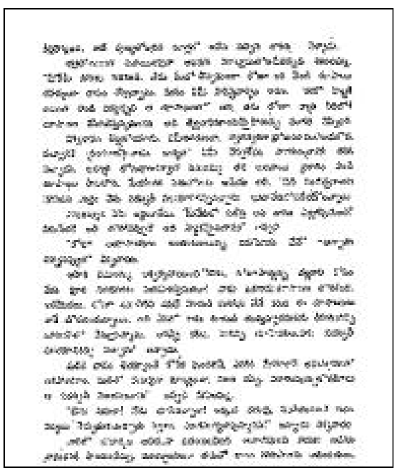

Figure 2 A Sample page in Telugu script

A set of distinct characters is identified as the training set. Signatures of each of the images are computed with different filter banks. The DNN is trained with the set of signatures of the training set. Once the network is trained, it is presented with the remaining set of characters from the scanned pages and the performance is noted. The recognition rate is computed as the ratio of correctly recognized input characters to the total number of test input characters. We observe that a very conservative recognition rate using Battle-Lemarie filter can be as close to 93.46%. The performance across fonts and sizes varied from 93% to 95%. The performance of the recognizer for different set of filters is reported in Table 1.

	B-L filter		A-S filter	
	Set 1	Set 2	Set 1	Set 2
No. of Training Patterns	52	52	52	52
No. of Testing Patterns	444	100	100	100
No. of Correct Recognitions	415	93	85	87

Table 1 Experimental Results

6. Conclusions

We presented a very comprehensive and practical OCR system for Telugu language. The proposed system is shape and font independent and does require any preprocessing. The signatures are computed using proper selection of Wavelet scaling function such that the basic structure of Telugu symbols are made use of. It is to be noted that we use the same system to recognize English character and observed that the recognition rate is very low. This is expected as the directional features which is prevalent in Latin scripts are not preserved during signature computation.

Figure 3 **Incorrectly recognized character. (a) Input (b) Output characters**

It is also interesting to note that some pairs of Telugu symbols normally mis-recognised. Two such pairs are shown in Figure 3. It may be seen that the two shapes of characters resemble each other with difference in the region at the bottom and at the top. To conclude, the design, approach and implementation were driven by the need for a practical OCR system for Telugu character recognition.

References

[Burr 88] D. J. Burr. Experiments on neural net recognition of spoken and written text, *IEEE Trans on ASSP* 36, No. 7, 1988.

[Casey, 66] R. Casey, G. Nagy. Recognition of printed Chinese character, IEEE tr Electron. Comp, EC-15 (1966) pp 91-101.

[Chaudhuri 91] B.B. Chaudhuri, O.A. Kumar and K.V.Ramana. Automatic generation and recognition of Telugu script characters, *J IETE* 37, 499-511,1991.

[Chaudhuri, 98] B. B.Chaudhuri and U. Pal. A complete Printed Bangla OCR System, Pattern Recognition, vol. 31, no. 5, pp. 531-549, 1998.

[Cheung, 2001] A Cheung, M Bennamoun, N. W. Bergmann, An Arabic OCR system using recognition-based segmentation, Pattern Recognition, Vol 34, pp 215-233, 2001.

[Devijver, 82] P. A. Devijver and J. Kittler, Pattern Recognition: A Statistical Approach. London: Prentice-Hall, 1982.

[Dunn, 92] C. E. Dunn and P. S. P. Wang, Character segmentation techniques for handwritten text --- a survey, in Proceedings of 11th IAPR Int. Conf. Pattern Recognition, vol. II, (The Hague, The Netherlands), pp. 577--580, IEEE Computer Society Press, 1992.

[Duda, 73] R. O. Duda and P. E. Hart, Pattern Classification and Scene Analysis. New York: John Wiley & Sons, 1973.

[Elliman, 90] D. G. Elliman and I. T. Lancaster. A review of segmentation and contextual analysis for text recognition, Pattern Recognition, vol. 23, no. ¾, pp. 337--346, 1990.

[Fu, 82] K.-S. Fu. Syntactic Pattern Recognition and Application. Englewood Cliffs, New Jersey: Prentice-Hall, 1982.

[Fukunaga, 90] K. Fukunaga. Introduction to Statistical Pattern Recognition. Boston: Academic Press, 1990.

[Fukushima, 91] K. Fukushima and N. Wake. Handwritten alphanumeric character recognition by the neocognitron, IEEE Transactions on Neural Networks, vol. 2, pp. 355--65, May 1991.

[Govindan, 90] V. K. Govindan and A. P. Shivaprasad. Character Recognition- A Review, Pattern Recognition, vol. 23, no. 7, pp. 671--683, 1990.

[Grother, 92] P. J. Grother. Karhunen Loeve Feature Extraction for Neural Handwritten Character Recognition, In Proceedings: Applications of Artificial Neural Networks III, Orlando, SPIE, April 1992.

[Guyon, 91] I. Guyon. Applications of Neural Networks to Character Recognition, In P.S.P. Wang (Ed), Character and Handwriting Recognition: Expanding Frontiers, 1991.

[Hubrig, 92] S. Hubrig-Schaumburg. Handwritten character recognition using a reflective modular neural network system, November 1992.

[Impedevo, 91] S. Impedevo, L Ottaviano and S. Occhingro. Optical Character Recognition - A Survey, International Journal of Pattern Recognition and Artificial Intelligence, 1991.

[Kim, 2001] H.- Y. Kim, J.H. Kim, Hierarchical random graph represetation of handwritten ans its application to Hangul recognition. Pattern Recognition, Vol 34, 2001.

[Kunihito, 92] F. Kunihito and W. Nobuaki. Handwritten alphanumeric character recognition by the neurocognitron. IEEE Trans. on Neural Networks, 2(3):355—365, 1992.

[Lam, 92] L. Lam, S.-W. Lee, and C. Y. Suen, Thinning methodologies --- a comprehensive survey, IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 14, pp. 869--885, Sept. 1992.

[Lovell. 94] D. R. Lovell, The Neocognitron as a system for handwritten character recognition: limitations and improvements. PhD dissertation, University of Queensland, July 1994.

[Mantas, 86] J. Mantas. An overview of character recognition methodologies, Pattern Recognition, vol. 19, no. 6, pp. 425--430, 1986.

[Mallat, 89] S.Mallat. Multiresolution Approximations and Wavelet Orthonormal Basesof L^2R, Transactions of the American Mathematical Society., vol3, No.1, pp 69-87, 1989

[Mori, 92] S. Mori, C. Y. Suen, and K. Yamamoto. Historical review of OCR research and development, Proceedings of the IEEE, vol. 80, pp. 1029--1058, July 1992.

[Mohiuddin, 94] K. Mohiuddin and J. Mao. A comparative study of different classifiers for handprinted character recognition. In E. S. Gelsema and L. N. Kanal, editors, Pattern Recognition in Practice IV, pages 437--448. Elsevier Science, The Netherlands, 1994.

[Negi, 2001] Atul Negi, Chakravarthy Bhagvati, B.Krishna, An OCR system for Telugu, 6^{th}ICDAR, Seattle, USA, Sept. 2001.

[Nagy, 92] G. Nagy, At the frontiers of OCR, Proceedings of the IEEE, vol. 80, pp. 1093--1100, July 1992.

[Perantonis, 92] S. J. Perantonis and P. J. G. Lisboa. Translation, rotation, and scale invariant pattern recognition by high-order neural networks and moment classifiers, IEEE Trans on Neural Net works 3, 241-251, 1992

[Pentland, 91] A. Pentland. Eqilibrium and Interpolation solutions using Wavelet Bases, Proceedings of CG, '91 Conference, pp 507-523, Newyork, Springer, 1991.

[Rajasekhar, 77] S. N. S. Rajasekharan and B. L. Deekshatulu. Generation and Recognition of printed Telugu characters, Computer graphics and image processing 6, 335 – 360, 1977.

[Rao, 99] M. S. Rao, Arun K Pujari. A New Neural Network architecture with associate memory, Pruning and order sensitive learning. International Journal of Neural Systems, 9, 4, 351-370, 1999.

[Sabourin, 92] M. Sabourin and A. Mitiche. Optical character recognition by a neural network, Neural Networks 5, 843—852, 1992

[Schwenk, 95] H. Schwenk and M. Milgram. Transformation invariant autoassociation with application to handwritten character recognition, In D.S. Touretzky, G. Tesauro, and T.K. Leen, (Eds.), Neural Information Processing Systems 7, (pp. 991-998, MIT Press, 1995.

[Sinha 79] R. M. K. Sinha and H. N. Mahabala. Machine recognition of Devanagari script, IEEE Trans on SMC –9, 8, 435 – 441, 1979.

[Sinha 87] R. M. K. Sinha. Rule based contextual post-processing for Devanagari text recognition, *Pattern Recognition* 20, 475-485,1987.

[Suen 80] C. Y. Suen, M. Berthod, and S. Mori. Automatic recognition of hand printed characters – The state of the art, *Proc of IEEE* 68, 469-487, 1980.

[Sukhaswami 95] M. B. Sukhaswami, P. Seetharamulu, Arun K Pujari. Recognition of Telugu characters using neural networks. Int. J. of Neural Systems, 6(3):317-357, 1995.

[Trier, 96] O. D. Trier, A K Jain, T. Taxt. Feature extraction methods for Character Recognition- A Survey Pattern Recognition, vol 29, pp 641-662, 1996.

[Ullmann, 73] J. R. Ullmann. Pattern Recognition Techniques. London, England: Butterworth, 1973.

[Wilson, 93] C. L. Wilson. Evaluation of Character Recognition Systems, In Neural Networks for Signal Processing III, IEEE, New York, pp. 485-496, September 1993.

[Wang, 2001] A.-B. Wang, K.-C. Fan. Optical recognition of handwritten Chineses character by hierarchical radical matching method, Pattern Recognition, 34, pp 15-35, 2001.