Machine Recognition of Printed Kannada Text

B. Vijay Kumar and A.G. Ramakrishnan

Department of Electrical Engineering,
Indian Institute of Science, Bangalore 560012, India
{vijaykb,ramkiag}@ee.iisc.ernet.in

Abstract. This paper presents the design of a full fledged OCR system for printed Kannada text. The machine recognition of Kannada characters is difficult due to similarity in the shapes of different characters, script complexity and non-uniqueness in the representation of diacritics. The document image is subject to line segmentation, word segmentation and zone detection. From the zonal information, base characters, vowel modifiers and consonant conjucts are separated. Knowledge based approach is employed for recognizing the base characters. Various features are employed for recognising the characters. These include the coefficients of the Discrete Cosine Transform, Discrete Wavelet Transform and Karhunen-Louve Transform. These features are fed to different classifiers. Structural features are used in the subsequent levels to discriminate confused characters. Use of structural features, increases recognition rate from 93% to 98%. Apart from the *classical* pattern classification technique of nearest neighbour, Artificial Neural Network (ANN) based classifiers like Back Propogation and Radial Basis Function (RBF) Networks have also been studied. The ANN classifiers are trained in supervised mode using the transform features. Highest recognition rate of 99% is obtained with RBF using second level approximation coefficients of Haar wavelets as the features on presegmented base characters.

1 Introduction

Kannada, the official language of the south Indian state of Karnataka, is spoken by about 48 million people. The basic structure of Kannada script is distinctly different from Roman script. Unlike many North Indian languages, Kannada characters don't have shirorekha (a line that connects all the characters of any word) and hence all the characters in a word are isolated. This creates a difficulty in word segmentation. Kannada script is more complicated than English due to the presence of compound characters. However, the concept of upper/lower case characters is absent in this script.

Modern Kannada has 48 base characters, called as *varnamale*. There are 14 vowels (Table 1) and 34 consonants. Consonants are further divided into grouped consonants (Table 2) and ungrouped consonants (Table 3). Consonants take modified shapes when added with vowels. Vowel modifiers can appear to the right, on the top or at the bottom of the base consonant. Table 4 shows the shapes of the consonant 'ಠ' when modified by vowels. Such consonant-vowel

D. Lopresti, J. Hu, and R. Kashi (Eds.): DAS 2002, LNCS 2423, pp. 37–48, 2002.

combinations are called modified characters. Same consonants combine to form consonant conjucts (Table 5). In addition, two, three or four characters can generate a new complex shape called a compound character.

Table 1. Vowels and their ASCII representations

ಅ	ಆ	ಇ	ಈ	ಉ	ಊ	ಋ
a	aa	e	ee	u	oo	Ru
ೠ	ಎ	ಏ	ಐ	ಒ	ಓ	ಔ
Roo	ae	aee	i	o	O	au

Table 2. Grouped consonants and their ASCII representations

ಕ	ಖ	ಗ	ಘ	ಙ
ka	kha	ga	gha	Gnya
ಚ	ಛ	ಜ	ಝ	ಞ
ca	cha	ja	jha	Jnya
ಟ	ಠ	ಡ	ಢ	ಣ
ta	Ta	Da	Dha	Na
ತ	ಟ	ದ	ಧ	ನ
tha	Tha	da	dha	na
ಪ	ಫ	ಬ	ಭ	ಮ
pa	pha	ba	bha	ma

Table 3. Ungrouped consonants and their ASCII representations

ಯ	ರ	ಲ	ವ	ಶ	ಷ	ಸ	ಹ	ಳ
ya	ra	la	va	Sa	Sha	sa	Ha	La

Table 4. Modification of base consonant by vowels

Vowel	Vowel Modifiers	When attached to consonat(ರ)
ಅ	◌	ರ
ಆ	◌ಾ	ರಾ
ಇ	◌ಿ	ರಿ
ಈ	◌ೀ	ರೀ
ಉ	◌ು	ರು
ಊ	◌ೂ	ರೂ
ಋ	◌ೃ	ರೃ
ೠ	◌ೄ	ರೄ
ಎ	◌ೆ	ರೆ
ಏ	◌ೇ	ರೇ
ಐ	◌ೈ	ರೈ
ಒ	◌ೊ	ರೊ
ಓ	◌ೋ	ರೋ
ಔ	◌ೌ	ರೌ

Table 5. Consonant conjuncts

ಕ	ಖ	ಗ	ಘ	ಙ	ಚ	ಛ	ಜ	ಝ	ಞ	ಟ	ಠ	ಡ	ಢ	ಣ	ತ	ಥ
ದ	ಧ	ನ	ಪ	ಫ	ಬ	ಭ	ಮ	ಯ	ರ	ಲ	ವ	ಶ	ಷ	ಸ	ಹ	ಳ

1.1 Line Segmentation

To segment the lines, the horizontal projection profile (HPP) of the document is obtained. HPP is the histogram of the number of ON pixels accumulated horizontally along every pixel row of the image. This profile has valleys of zero height between the lines, which serve as the separators of the text lines as depicted in Fig. 1. In the case of Kannada script, sometimes the bottom conjuncts of a line overlap with the top-matras of the following text line in the projection profile. This results in non-zero valleys in the HPP as shown in Fig. 2. These lines are called *kerned* [2] text lines. To segment such lines, the statistics of the heights of

the lines are found out from the HPP. Then the threshold is fixed at 1.6 times the average line height. This threshold is chosen based on experimentation of our segmentation algorithm on a large number of Kannada documents. Non-zero valleys below the threshold indicate the locations of the text line and those above the threshold correspond to the location of *kerned* text lines. The mid point of a non-zero valley of a *kerned* text line is the separator of the line.

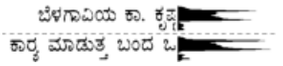

Fig. 1. The dotted lines indicate the obtained line boundaries.

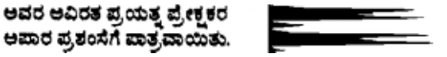

Fig. 2. Kerned text lines. The horizontal projection profile does not have zero valleys in between the text lines.

Fig. 3. Word segmentation: A. Input text line, B. Text line image after dilation. C. Vertical projection of image in B. The zero valleys in the projection separate the words.

1.2 Word Segmentation

Kannada words do not have *shirorekha*, all the characters in a word are isolated. Further, the character spacing is non-uniform due to the presence of consonant conjuncts. In fact, whenever the latter are present, the spacing between the base characters in the middle zone becomes comparable to word spacing. This could affect the accuracy of word segmentation. Hence, morphological dilation [3] is used to connect all the characters in a word, before performing word segmentation. Each ON pixel in the original image is dilated with a structuring element. Based on experimentation, we found that, for a scanning resolution of 400 DPI, a structuring element of size 2x6 with all 1's (foreground) is adequate to connect all the characters in a word. Then, the vertical projection profile (VPP) of the dilated image is determined. This is the histogram of column-wise sum of ON pixels. The zero-valued valleys in the profile of the dilated image separate the words in the original image. This is illustrated in Fig. 3.

1.3 Zone Detection

Based on the HPP, each word is partitioned into three zones as depicted in Fig. 4. The imaginary horizontal line passing through the index corresponding to the maximum in the top half of the profile is the headline (starting of the ascenders) and the baseline corresponds to the maximum in the bottom half of the profile (starting of the descenders). The top zone denotes the portion above the headline, where top matras or ascenders occur. The gap between headline and baseline is the middle zone, which covers base and compound characters. Portion below the baseline is bottom zone in which bottom matras or descenders of aspirated characters occur. Also, the consonant conjuncts (see Table 5) occur in this zone.

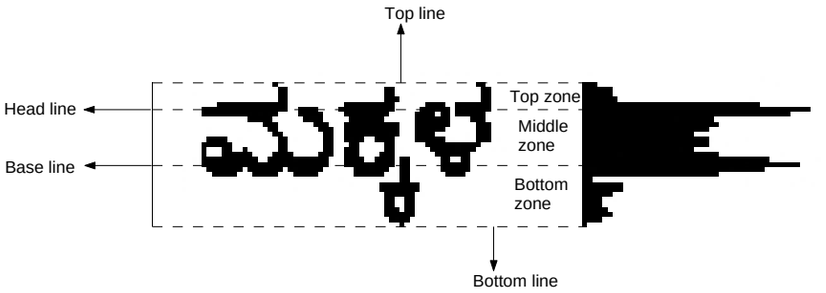

Fig. 4. Different zones of a Kannada word

1.4 Character Segmentation

Zone detection helps in character segmentation. Adjacent characters in a Kannada word sometimes overlap in the VPP due to the presence of consonant conjuncts as shown in Fig. 5(b). These are called *kerned* characters. Such characters cannot be segmented using zero-valued valleys in the projection profile. Using the baseline information, the text region in the *middle* and *top* zones of a word is extracted and its VPP is obtained. Zero-valued valleys of this profile are the separators of the characters (see Fig. 5(d)). Sometimes, the part of a consonant conjunct in the middle zone is segmented as a separate symbol. Such things are eliminated in the recognition phase, based on the total number of ON pixels in the symbol. The total number of Kannada characters, including base characters, characters formed with consonant-vowel combination, consonants with conjuncts and compound characters, are 34*34*14 + 14. This results in a huge number of classes for recognition, which is difficult to handle. So, we split the segmented character into a *base character* and a *vowel modifier* (top or right matra). The consonant conjuncts are segmented separately based on connected component analysis (CCA).

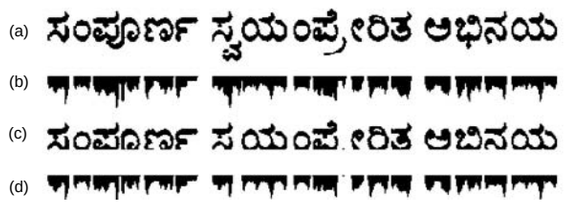

Fig. 5. Character segmentation. (a) Input text image. (b) Its vertical projection. (c) The text part of the image (a) in middle and top zone above the baseline. (d) Vertical projection of image in (c). The zero valleys in this profile separate the characters.

Consonant Conjunct Segmentation. *Knowledge* based approach is used to separate the consonant conjuncts. The spacing to the next character in the middle zone is more for characters having consonant conjuncts than it is for others. To detect the presence of conjuncts, a block of image in the *bottom zone* corresponding to the gap between adjacent characters in the *middle zone* is considered as shown in Fig. 6. We call this image block as *partial image*. If the number of ON pixels in the partial image exceeds a threshold (set 15 pixels), a consonant conjunct is detected. Sometimes a part of the conjunct enters the middle zone between the adjacent characters. Such parts will be lost if the conjunct is segmented only in the bottom zone. Thus, in order to extract the entire conjunct, we use CCA. However, in some cases, the conjunct is connected to the character in the middle zone, causing difficulty in using CCA for segmenting only the conjunct. To address this problem, the character in the middle zone is removed before applying CCA. For example, in Fig. 6, CCA is applied on the image PQRS, after setting all the pixels in the part PMNO to zero. This results in the image shown in Fig. 7, which leads to the detection of three distinct connected components. The component with the maximum number of ON pixels is the conjunct.

Vowel Modifier Segmentation. This is divided into segmentation of *top* and *right* matras. The part of the character above the headline in the top zone is the *top matra*. Since the headline and baseline of each character are known and if the aspect ratio of the segmented character in the combined top and middle zones is more than 0.95, then it is checked for the presence of the right matra.

For *right matra* segmentation, three subimages of the character are considered as shown in Fig. 8: the whole character, head and tail images. The head image is the segment containing five rows of pixels starting from the headline downwards. Similarly, the tail image contains 5 rows downwards from the baseline. VPP for each of these images is determined. Let the index corresponding to the maximum of the profile of the character image be P. Let b1 and b2 be the indices corresponding to the first zero-values immediately after the index P, in the profiles of head and tail images, respectively. The break point is selected

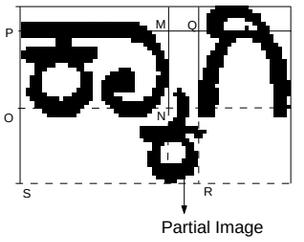

Fig. 6. Segmentation of consonant conjuncts

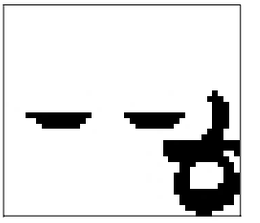

Fig. 7. Image used for connected component analysis, after setting the PMNO part of PQRS of Fig. 6 to background.

as the smaller of b1 and b2. The characters are normalized before feature extraction to avoid the effects of variations on the character image such as size and displacement. A separate normalization size is used for the base character and the vowel modifiers. The base characters are normalized to 32X32, while modifiers and consonant conjuncts are resized to a size of 16X16 using bilinear interpolation.

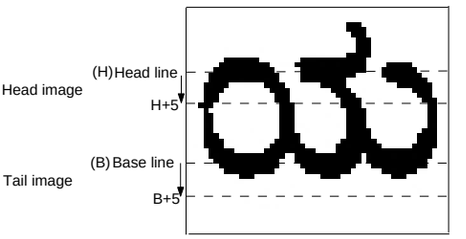

Fig. 8. Different subimages of a character considered for vowel modifier segmentation

2 Results

Fig. 9 gives the flow chart of the total Kannada OCR system. The software is implemented on a Sun-Ultra Sparc workstation using C language and $MATLAB$ package under Unix platform. The data samples are collected by scanning various Kannada magazines with a resolution of 400 DPI. More than 40,000 characters are collected from these images. The results presented are based on presegmented characters. The training set for the base characters contains 1110 samples corresponding to 37 classes and each category has 30 patterns. In order to make the system more robust, some noisy characters are also included in the training

set. The performance of various features and classifiers have been evaluated on a test set containing 1453 randomly selected characters with different font styles and sizes. The results corresponding to different techniques employed for feature extraction and classification are presented in the subsequent sections and these results are based on presegmented characters. The vowels in Kannada script occur only at the beginning of a word. This information, if used while testing, not only improves recognition accuracy but also helps in fast classification. Since this *knowledge* is used in our work, each pattern in the test set also contains the information about its position in the word. The training sets for the modifiers and the consonant conjuncts contain 180 and 540 patterns, corresponding to 9 and 27 classes, respectively. Various transform based features are used to evaluate the performance of nearest neighbour (NN) and Neural network based classifiers. The employed features include, the coefficients of the Discrete Cosine Transform (DCT), Discrete Wavelet Transform (DWT) and Karhunen-Louve Transform (KLT). The transforms are applied on the complete pattern image, rather than on the subblocks. Due to energy compaction property of DCT, only the significant coefficients are considered for recognition. However, in case of DWT, the approximate coefficients in the second level of decomposition are considered for recognition. The KLT transformation matrix is obtained from the training samples. Using the eigen vectors corresponding to different numbers of significant eigen values, results are obtained. Table 6 shows the recognition accuracy of NN classifier for base characters using the various features.

Table 6. Recognition accuracy of NN Classifier on various features

Feature block size	Recognition rate (%) without structural features	Recognition time (min)	Recognition rate (%) with structural features	Recognition time (min)
Discrete cosine transform				
4x4	91.81	0.48	93.80	0.60
8x8	93.54	1.03	98.70	1.23
12x12	92.63	2.00	98.27	2.11
Karhunen-Louve transform				
40	92.56	0.78	98.70	0.92
50	92.70	0.93	98.55	1.13
60	92.84	1.18	98.77	1.32
Discrete wavelet transform				
Haar (8x8)	92.42	1.83	98.83	2.65
db2 (10x10)	92.36	2.25	98.55	3.53

Table 7 lists the pairs of confused characters using the NN classifier on DCT features. DWT and KLT also gave almost the same confusion character pairs as DCT using the NN classifier. The recognition rate is improves by around 6%, on using the structural features in the second and third level to resolve the confused characters. The structural features, such as aspect ratio, orientation of particular

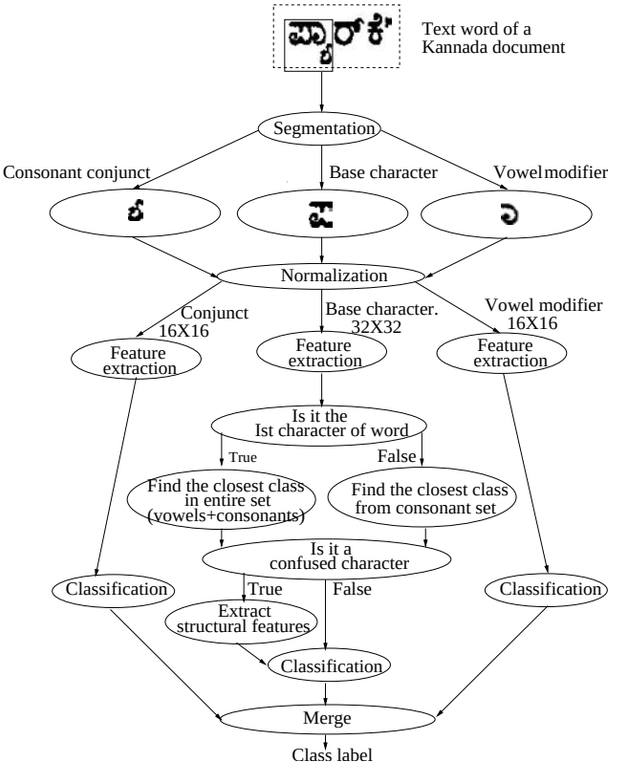

Fig. 9. Flow chart of the complete Kannada recognition system.

Table 7. Confused base character pairs using NN classifier on DCT features

ಬ	ಭ
ಭ	ಬ
ರ	ನ
ಪ	ಷ
ಭ	ಷ
ಆ	ಅ
ಭ	ಲ
ಖ	ಜ
ರ	ಂ

strokes, width of the middle zone and height of the segment in the top zone are extracted from the subimages of the character. Table 8 lists a group of confused characters and the structural features used to resolve them.

2.1 Performance Evaluation of Back Propagation Network (BPN) and Radial Basis Function (RBF) Network on Base Characters

The features found to be best in the previous sections are used to evaluate the performance of BPN and RBF networks [4]. The BPN was trained in batch mode using supervised learning employing logsigmoidal activation function. In order to obtain good generalization with the network, the weights and biases of the network are set to small random values. Because of the form of the activation function, the input is normalized to a range of 0 to 1 before training.

Table 8. Resolving confused characters using structural features

Confusion character set	ಪ ನ ಪ ಪ
ಪ	The number of ON pixels more than 40 in the orientation 40-70 degrees in the lower right quarter image
ನ	The length of maximum in the bottom half of image less than the 75% of the width of the character
ಪ	The number of ON pixels more than 35 in the orientation 20-50 degrees in the
ಪ	upper middle region of the image If all the above conditions are not satisfied

Also, the presentation of the training samples to the network is randomized. The RBF network employing Gaussian kernel as the activation function was trained in supervised mode of learning. The radial basis functions are centered on each training pattern and the layer biases are all kept constant depending on the spread of the Gaussian. The recognition performance studied for different values of the variance of the gaussian. In both the cases (BPN and RBF), structural features have been used in further levels of classification to discriminate between similar characters from different classes. The results corresponding to BPN and RBF are listed in the Tables 9 and 10, respectively. RBF performed

Table 9. Recognition accuracy using BPN with Haar (db1) wavelet features. [L1 and L2 are the number of nodes in the first and second hidden layers, respectively.]

Number of hidden layers	Number of hidden neurons	Recognition rate (%)	Recognition time (min)
Using 8x8 Haar features			
1	20	96.14	1.78
1	25	96.28	1.78
2	L1=35,L2=25	97.04	1.87
Using 8x8 DCT features			
1	20	95.87	1.81
1	25	95.73	1.64
2	L1=40,L2=25	94.15	0.98

better than the NN classifier and BPN with the same set of features. However, the highest recognition rate of around 99 % is achieved with Haar wavelets. On the other hand, DCT and KLT gave recognition accuracies of 98.8 % and 98.6 %, respectively with a spread of 11. As before, the performance using Haar features is consistently better. Advantage of the RBF network over the BPN is that, training time is very less.

Table 10. Recognition accuracy using RBF network with various features

Spread of Gaussian	Haar		DCT		KLT	
	Rec rate (%)	Rec time (min)	Rec rate (%)	Rec time (min)	Rec rate (%)	Rec time (min)
4	69.23	3.71	52.92	2.94	29.86	1.70
8	98.07	2.65	97.66	2.58	98.48	1.79
10	98.83	2.65	98.83	2.50	98.62	1.74
11	99.03	2.59	98.89	2.55	98.62	1.91

2.2 Recognition of Top and Right Matras

The training set for top and right matras, contains 9 classes with 20 samples in each class and the test set contains 345 patterns. NN and RBF network are compared for their performance. The spread of RBF is set based on the previous experimental results. The results are listed in Table 11. RBF performed better than NN classifier with the same set of features. However, the NN classifier classifies faster than RBF.

2.3 Recognition of Consonant Conjuncts

The training set contains 27 classes with 20 samples in each class and the test has 531 patterns. All the tests are performed with the 64-dimensional feature vector, which is found to be best in the previous sections. The BPN and RBF are trained in supervised mode of learning. In the case of wavelets, the approximation coefficients of the first level of decomposition are used as features. Results are shown in Table 12. Recognition performance of the NN classifier is better than those of RBF and BPN networks, employing the best features. The recognition time using db1 is always more than the time using DCT features. RBF and BPN might perform better than NN classifier if the various parameters of the network are properly tuned.

Table 11. Performance of NN classifier and RBF network on matras

Feature	Size of feature vector	NN Classifier		RBF, Spread=10	
		Recognition rate (%)	Recognition time (min)	Recognition rate (%)	Recognition time (min)
Haar (db1)	64	94.20	0.56	96.81	0.66
DCT	64	93.04	0.43	96.81	0.55

Table 12. Recognition of consonant conjuncts by various classifiers with Haar and DCT as features

Classifier	Size of feature vector	db1 (Haar)		DCT	
		Recognition rate (%)	Recognition time (min)	Recognition rate (%)	Recognition time (min)
NN	64	96.61	0.37	96.79	0.21
BPN (50)	64	95.10	0.40	93.78	0.24
RBF (S=10)	64	95.66	0.48	95.48	0.31

2.4 Final Recognition Results

The classifier outputs the labels corresponding to the recognized base character, vowel modifier and consonant conjunct in the middle, top/middle and bottom zones, respectively. The recognized modifier and consonant labels are then appropriately attached to the recognized base character label to produce the final character label. These labels are then mapped to customized codes and stored in a file. This file can be viewed using any compatible Kannada type-setting software. KanTex is [5] is one such software which is compatible with LaTeX type-setting tools. Fig. 10 shows a test document (top part of the image) with a large number of noisy characters and the corresponding recognized output (bottom) is also shown in Fig. 10. The symbol '*' indicates rejected characters.

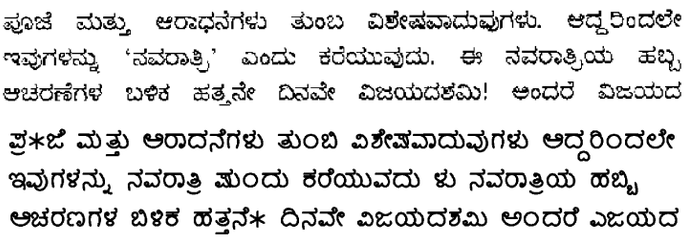

Fig. 10. Input test document (top) and recognized output

3 Conclusions

The present work addresses the issues involved in designing a full fledged OCR system for printed Kannada text. Recognition of Kannada characters is more difficult than many other Indian scripts due to higher similarity in character shapes, a larger set of symbols and higher variability across fonts in the characters belonging to the same class. The performance evaluation of the various classifiers using transform based features has been presented. Experimental results show that employing structural features in the second stage of classification improves

the recognition accuracy. This kind of hierarchical classification makes a high recognition rate possible with a small dimensional set of features. In the case of base characters, the performance of the RBF networks using Haar wavelet features followed by the structural features resulted in the highest recognition rate of around 99.03 %. On the other hand, NN classifier and BPN using the same input features achieved recognition rates of 98.83 % and 97.04 %, respectively. However, the recognition time of NN classifier was less than those of RBF and BPN.

In the case of consonant conjuncts, NN classifier performs better than RBF and BP networks. The recognition rate using NN classifier is 96.79 % with a recognition time of 0.21 minute by selecting the best of the features. On the other hand, RBF and BPN achieved recognition rates of 95.66 % and 95.10 % with recognition times of 0.25 min and 0.31 min, respectively.

For the recognition of the top and right matras, two different classifiers are applied on the test set containing 345 characters with 64-dimensional best feature vector. The recognition performance of RBF is 96.8 % and that of NN classifier is 94.2 % employing as features the approximation coefficients of the Haar wavelets in the first level of decomposition.

References

1. Chaudhuri, B.B., Pal, U.: A Complete Printed Bangla OCR System. Pattern Recognition, Vol. 31, No. 5 (1998) 531–549
2. Lu, Y.I.: Machine Printed Character Segmentation – An Overview. Pattern Recognition, Vol. 28, No. 1 (1995) 67–80
3. Gonzalez, R.C., Woods, R.E.: Digital Image Processing. Addison Wesley, New York (1993)
4. Haykin, S.: Neural Networks. A Comprehensive Foundation. Pearson Education Asia (1999)
5. Jagadeesh, G.S., Gopinath, V.: Kantex, A Transliteration Package for Kannada. Kantex Manual. http://langmuir.eecs.berkeley.edu/venkates/KanTex_1.00.html