Feature Extraction (Image Compression) of Printed Gujarati and Amharic Letters Using Discrete Wavelet Transform

Archit Yajnik[1]* and Dharmendra Singh[2]

ABSTRACT

This paper demonstrates an application of discrete wavelet transform as a feature extractor as applicable to natural language processing. The procedure discussed in this paper extracts important features of the printed characters of Gujarati and Amharic scripts and shows Image compression capabilities of discrete wavelets. Procedure prescribed in this paper compresses the original image to 75% without disturbing much the original image.

Keywords : *Amharic script, Discrete Wavelet Transform (DWT), Daubechies D4, Feature extraction, Image Compression, $l^2(Z_N)$-Space of Square summable sequences of N Complex number.*

I. INTRODUCTION

THIS paper deals with two different scripts viz. Gujarati (Gujarati language, India) and Amharic (Ge'ez letters, Ethiopia). Both the scripts are entirely different as far as the shapes and curvature are concerned. Development of optical character recognition (OCR) for various scripts is an upcoming field of research [2, 4, 5, 6, 7]. This task is divided into several subtasks. Among those feature extraction plays a prominent role in the development of OCR. Various image compression and feature extraction techniques like statistical moments, fringes, Fourier descriptors, wavelets etc have been used for different Indic and European scripts in the literature. But the same is not explored in the case of both the scripts. This paper presents an approach of Wavelet descriptors for the extraction of important features of the printed Amharic & Gujarati symbols. Daubechies D4 Wavelet coefficients are used for this purpose.

An overview of the paper is as follows: In section-2, character modeling of the Gujarati and Amharic numerals is discussed, a brief review of the Daubechies Wavelets and feature extraction technique are introduced in the Section-3. Section 4 demonstrates experimental procedure and results, followed by conclusion in Section 5.

II. CHARACTER MODELING

A. Character Images

The representations of characters for both the languages are depicted in table 2.1. In this paper we have used the images of numeral letters from Power Ge'ez '06 for Amharic language.

Table 2.1
Character modeling for Gujarati and Amharic numerals

Font description	**Character Images**
Gujarati (normal) 15	૦ ૧ ૨ ૩ ૪ ૫ ૬ ૭ ૮ ૯
Amharic (normal) 11	

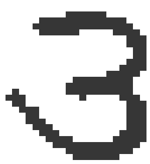
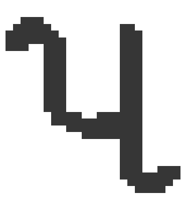
B. Scanning and binarization

The character images were selected from binarized images of documents scanned at 300 pixels per inch resolution using the HP-Scan Jet II scanner. The character images were normalized to (32 x 32) array of binary pixels. The Daubechies D4 wavelet transform was applied to the character images and 256 low-low coefficients were used to construct the feature

1*Archit Yajnik is Reader with the Department of Humanities and Applied Sciences, SMS Institute of Technology, Lucknow.
2 Dharmendra Singh is Reader with the Department of Humanities and Applied Sciences, SMS Institute of Technology, Lucknow.

vector. Table 2.1 depicts the scanned and binarized images of Gujarati /Three and /Five.

Table 2.1 : Scanned and Binarized Representations of Gujarati Numerals

Scanned Image	Binarized image
	11111111111111000000000111111111 11111111111111000000000111111111 11111111000000000000000000001111 11111110000000000000000000000111 11111111000000000111111110000001 11111111000000000111111110000001 11111111111111111111111111110000 11111111111111111111111111110000 11111111111111111111111111111000 11111111111111111111111111111000 11111111111111111111111111111000 11111111111111111111111111110000 11111111111111111111111111000000 11111111111111111111111000000001 11111111111111111111111000000001 11111111111111110000000000000111 11111111111111000000000000000111 11001111111111000000000000000111 00000111111111111001111111000001 00000111111111111001111111000001 11000111111111111111111111110000 11110000111111111111111111110000 11111000111111111111111111110000 11111000111111111111111111110000 11111000001111111111111111110000 11111110000111111111111111110000 11111111000001111111111111000000 11111111110000001111111110000000 11111111110000001111111110000000 11111111111000000000000000000001 11111111111111000000000000000111 11111111111111110000000000111111
	11100001111111111111111111111111 11100001111111111111111111111111 11000000011111111111100011111111 00000000001111111111100000111111 00000000000011111111100000111111 00000110000011111111100000111111 00000110000011111111100000111111 11111110000011111111100000111111 11111110000011111111100000111111 11111110000011111111100000111111 11111110000011111111100000111111 11111110000011111111100000111111 11111110000011111111100000111111 11111110000011111111100000111111 11111110000011111111100000111111 11111110000011111111100000111111 11111110000011111111100000111111 11111110000011111111100000111111 11111111100000111000000000111111 11111111100000000000000000111111 11111111111100000000000000111111 11111111111111000000000000111111 11111111111111000000000000111111 11111111111111111111100000111111 11111111111111111111100000111111 11111111111111111111100000111111 11111111111111111111100000111111 11111111111111111111100000111111 11111111111111111111100000110000 11111111111111111111100000000000 11111111111111111111111000000001 11111111111111111111111100000011

III. DISCRETE WAVELETS AND FEATURE EXTRACTION TECHNIQUE

A. Discrete Wavelets

A spatially localized basis is useful in signal processing because it provides a local analysis of a signal. Let *z(n)* represent a signal with frequency *n*. If a certain coefficient in a series representation of *z* is large, we can identify the location with which this large coefficient is associated if the coefficients are spatially localized. We could then, for example, focus on this location and analyze it in greater detail.

Discrete Fourier transform [3] of a function $z \in l^2(Z_N)$ is given by

$$\hat{Z}(m) = \sum_{n=0}^{N-1} z(n) e^{-2\Pi imn/N}$$

The Fourier basis is not frequency localized, since the terms for any given frequency *n* in $\hat{Z}(m)$ will have magnitude 1. Development of discrete wavelets is based on scaling function which can briefly be defined as below:

Scaling functions are considered as the set of expansion functions composed of integer translations and binary scaling of the real, square-integrable function $\phi(x)$; that is, the set $\left\{\phi_{j,k}(x)\right\}$ where

$$\phi_{j,k}(x) = 2^{\frac{j}{2}} \phi(2^j x - k) \text{ for all } j,k \in \text{ Z.} \qquad (2.1)$$

Here, *k* determines the position of $\phi_{j,k}(x)$ along the x-axis, j determines $\phi_{j,k}(x)$'s width - how broad or narrow it is along the x-axis and the term $2^{j/2}$ controls its height or amplitude [3]. Because the shape of $\phi_{j,k}(x)$ changes with *j*, $\phi(x)$ is called a scaling function. By choosing $\phi(x)$ wisely, $\{\phi_{j,k}(x)\}$ can be made to span $L^2(R)$, the set of all measurable, square-integrable functions.

If we restrict *j* to a specific value, say $j = j_0$, the resulting expansion set $\{\phi_{j_0,k}(x)\}$, is a subset of $\{\phi_{j,k}(x)\}$. It will not span $L^2(R)$, but a subspace within it. The subspace can be defined as $V_{j_0} = span \{\phi_{j_0,k}(x)\}$.

That is, V_{j_0} is the span of $\{\phi_{j_0,k}(x)\}$ over k. If $f(x) \in V_{j_0}$, it can be written as

$$f(x) = \sum_k u_k \phi_{j_0,k}(x)$$

More generally, we will denote the subspace spanned over k for any j as

$$V_j = span\left\{\phi_{j,k}(x)\right\}$$

Following example demonstrates the Haar scaling function $\phi : R \to R$ Let be defined by

$$\phi(x) = \begin{cases} 1 & x \in [0,1) \\ 0 & otherwise \end{cases}$$

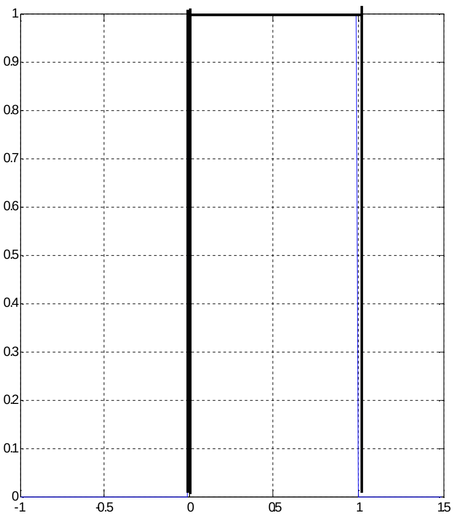


Fig. 2.1 : $f_{0,0}$ (x)

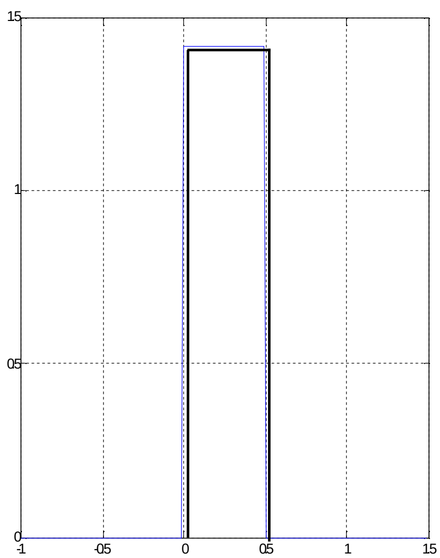


Fig. 2.2 : 1,0

Define $\phi_{j,k} : R \to R$ as

$$\phi_{j,k}(x) = 2^{\frac{j}{2}} \phi(2^j x - k)$$

Here, $\phi_{j,k}$ is known as the father wavelet. The pictorial representations of $\phi_{0,o}$ and $\phi_{1,o}$ are shown in the figures 2.1 and 2.2 respectively.

Define the vector space V^j as

$$V^j = span\left\{ \phi_{j,k} \right\}_{j,k \in Z}$$

Therefore V^j can be expressed as a linear combination of $\phi_{j,k}$ as follows:

$$V^j = \left\{ \sum_k u_j(k) \phi_{j,k} ; u = \left(u(k)\right)_{k \in Z} \in l^2(Z) \right\}$$

where $u_j(k)$ is called approximation coefficients at level *j*.

Since $\phi \in V^0 \subseteq V^1$, the above expression implies that

$$\phi(x) = \sum_{k \in Z} u_1(k) \phi_{1,k}(x)$$

$$= \sum_{k \in Z} u_1(k) \sqrt{2}\, \phi(2x - k)$$

and hence is called scaling function.

Let V^1 be an inner product space and let V^0 be subspace of the space V^1. Let W^0 be an orthogonal complement of V^0 in V^1 so that W^0 is also a subspace of V^1 . Hence $V^0 \oplus W^0 = \{ v_0 + w_0, v_0 \in V^0 \, and \; w_0 \in W^0 \}$ is called orthogonal direct sum of V^0 and W^0.

In particular, if we say $V^0 \oplus W^0 = V^1$, we mean that V^0 and W^0 are subspaces of V^1, $V^0 \oplus W^0$ and every element of $x \in V^1$ can be written as $x = u + v$ for some $u \in v^0$ and $v \in W^0$

$$\therefore \; V^1 = V^0 \oplus W^0$$

i.e. let $f(x) \in L^2(R)$ be a function defined in V^1. Therefore, it's wavelet expansion using wavelet function $\psi(x)$ (equation 2.1) and scaling function $\phi(x)$ (equation 2.1) is given by

$$f(x) = \phi_{1,k}(x) = \sum_k u_0(k) \phi_{0,k}(x) \quad + \quad \sum_k v_0(k) \psi_{0,k}(x) \qquad (2.2)$$

where the coefficients u_0 and v_0 are known as approximation and detail coefficients respectively.

Using equation (2.2) multistage wavelets can be developed easily (3). The most popular discrete wavelets have been developed by Daubechies which can be briefly discussed below:

Ingrid Daubechies constructed families of basis vectors called wavelets that are very well localized in space as well as in frequency for $l^2(Z_N)$. In the case of wavelets, instead of looking for an expansion of single vector z whose full set of translates form an orthonormal basis, we look for two vectors u and v such that the set of their translates by even integers forms an orthonormal basis. For example, in the case of Daubechies D4 wavelets, there are only four non-zero coefficients in the expansions of u and v, ie, they have representations of the following form:

$$u = \frac{\sqrt{2}}{8}\left(1+\sqrt{3},3+\sqrt{3},3-\sqrt{3},1-\sqrt{3},0,0,0,\cdots,0\right)$$

$$v = \frac{\sqrt{2}}{8}\left(-3-\sqrt{3},1+\sqrt{3},0,0,0,\cdots,-1+\sqrt{3},3-\sqrt{3}\right)$$

u and *v* are called the father wavelet and the mother wavelet respectively.

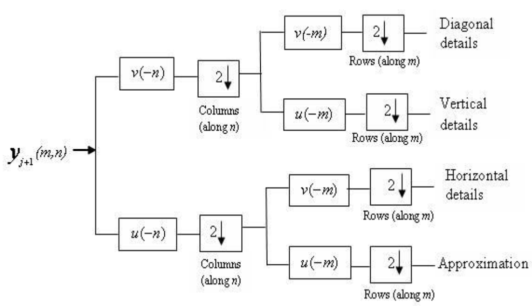


Fig. 3.1 Image Decomposition using DWT

B. Wavelet transform as a feature extractor

Like the one-dimensional discrete wavelet transform, the two-dimensional discrete wavelet transform can be implemented using digital filters and downsamplers. As in the one-dimensional case, image *f (x, y)* is used as the y_{j+1} *(m, n)* input. Convolving its rows with *v(-n)* and *u(-n)* and downsampling its columns, we get two subimages whose horizontal resolutions are reduced by a factor of 2. The highpass or detail component characterizes the image's high-frequency information with vertical orientation. Both subimages are then filtered columnwise and downsampled to yield four quarter-size output subimages viz. diagonal (High-High), vertical (High-Low), horizontal (Low-High) detail components and one approximation components (Low-Low). This process is known as first level decomposition. These images are the inner products of *f (x, y)* and the two-dimensional scaling and wavelet functions (shown as mother and father wavelet coefficients), followed by downsampling by two in each dimension.

Figure(3.1) shows the process of computing Discrete Wavelet Transform of a function *f(x,y)* of size *M x N*. Let $y_{j+1}(m,n)$ be the $(j+1)^{st}$ level of resolution of f where $0 \le m < M$ and $0 \le n < N$. The process starts by convolving y_{j+1} with high pass or detail components *v* (column wise, along *n*) and then it is down sampled. Same process is repeated for low pass or approximation components, *u*. The resultant outputs are further convolved with *v* and *u* respectively (row wise, along m) and down sampled. These computations decomposes the given two dimensional array into four parts, viz. diagonal (High-High), vertical (High-Low), horizontal (Low-High) detail components and one approximation components (Low-Low). That is known as first level decomposition.

In our experiments of character recognition of Gujarati script, the procedure of feature extraction of a normalized image of the size 32 x 32 can be summarized as below:

1. Input : The image is binarized in to 32 x 32 matrix. This binarized matrix is given as an input to the algorithm.
2. Convolution: Convolve this binarized image with father wavelet v and mother wavelet u followed by down- sampling by a factor of 2 twice. The image is now divided in to four subbands as discussed in the next step.
3. The four subbands (2^2) parts viz. low-low, low-high, high-low, high-high coefficients as shown in figure 3. Each part is of the size (16 x 16).
4. Consider only low-low (approximation) coefficients which capture the core information of the image.

These coefficients are considered as an input to the recognizer (like nearest neighborhood or Neural Network architectures [9, 10, 11]).

IV. EXPERIMENTAL PROCEDURE AND RESULTS

In the case of image processing, the input vector is made up of pixel intensities in an image. In our character images, we take an image of 32 x 32 pixels and apply convolution of the input data with the father wavelet u followed by down-sampling of the resultant data by a factor of 2. This process is repeated until the desired level of compression of the image data is achieved. In the present case, compression up to one level is implemented resulting in an output vector of size 16 x 16 pixels. The following pictures illustrate the original and the wavelet-compressed images for one and two of Gujarati (Table 4.1) alphabets /ma & /sha and Amharic (Table 4.2) numerals /one & /two. These are pixel maps with o's standing for black and spaces standing for white pixels.

Table 3 : Alphabets /ma & /sha

Image of Gujarati letter (Gujarati language)	32x32 Original binarized image	16x16 wavelet-compressed image (75%)
	11000000001111111111111111111111 11000000001111111111111111111111 00000000001111111111111111100111 00000000000011111111111111100000 00000000000000111111111111000000 00000000000000011111111111000000 11100000000000001111111111000001 11111111111000000011111111000001 11111111111100000011111111000001 11111111111100000011111111000011 11111111111000000011111100000011 11111110000000000000000000000011 11100000000000000010000100000011 11000000000000000000000000000011 00000000000000000000000000000011 00000000000000000000000000000011 00000001111000000000000000000011 00000001111000000000000000000011 00000011111000000011111100000011 00000011111000001111111100000011 00000000110000001111111100000011 11000001000000001111111100000011 11000000000000001111111100000011 11100000000000011111111100000011 11110000000000011111111100000001 11111100000001111111111100000001 11111111111111111111111100000001 11111111111111111111111100000001 11111111111111111111111111000001 11111111111111111111111111000000 11111111111111111111111111100000 11111111111111111111111111110001	1111111111111100 1000011111111101 0000011111111101 0000000111111000 1100000011111000 1111110001111000 1111110001110001 1100000000000001 0000000000000001 0000110000000001 0000110001110001 0000110011110001 1000000011110001 1100000011110000 1111111111110000 1111111111111000
	00000000111111111111111111100111 00000000111111111111111111100111 00000000001111111111111100000001 00000000000111111111111000000000 00000000000011111111110000000000 11111000000000111111111000000000 11111000000000111111111000000000 11111000000000011111111000000001 11111000000000001111111000000001 11111000000000000111111000000001 11111000000000000011111000000001 11111000000000000011111000000001 11111000000000000001111000000001 11111000000000100000011000000001 11111000000000110000001100000001 11111000000000111000000000000001 11111000000000111100000000000001 11111000000000111100000000000001 11111000000000111110000000000001 11111000000000111110000000000001 11111100000000011000000000000001 11111111000000000000000000000001 11111111000000000000000000000001 11111111110000000000000000000001 11111111111100000000000000000001 11111111111111000000000000000001 11111111111111110000011000000000 11111111111111110000011000000000 11111111111111111111110000000000 11111111111111111111110000000000 11111111111111111111111000000001	1111111111110000 0000111111111101 0000011111111101 0000001111110000 1110000111110000 1110000011110000 1110000001110000 1110000000110000 1110000100010000 1110000110000000 1110000111000000 1110000011000000 1111000000000000 1111110000000000 1111111100010000 1111111111110000

Table 4: Amharic Numerals / one & /two

Image of Ge'ez letter (Amharic Language)	32x32 Original binarized image	16x16 wavelet-compressed image (75%)
	11111111111111111111111111111111 11111111111111111111111111100111 11111111111100000000000000000111 11111111111100000111111000000111 11111111111100000111111110000111 11111111111100111111111111111111 11111111111111111111111111111111 11111111111111000000011111111111 11111111111111000000011111111111 11111111111111000000011111111111 11111111111111000000011111111111 11111111111111000000011111111111 11111111111100000000000111111111 11111111111100000000000001111111 11111111110000000001100000000111 11111111000000000111100000000111 11111111000000000111100000000111 11111111000000001111100000000111 11111111000000001111100000000111 11111100000000001111100000000111 11111111000000001110000000011111 11111111000000001110000000011111 11111111000000000000000001111111 11111111110000000000000111111111 11111111000000000000000000011111 11111111000000111110000000011111 11111100000011111111100000011111 11111111000011111111111111111111 11111111111111111111111111111111 11111111000011111100000011111111 11110000000011111100000011111111 11110000001111111100000011111111 11110000001111111111111111111111 11111111110000000000111111111111 11111111000000000000001111111111	1111111111111111 1111000111110011 1111110001111001 1111111111111111 1111111000011111 1111111000011111 1111110000001111 1111100000100001 1111000011000001 1111000011100001 1111000011000011 1111000000000111 1111000000000111 1111001111100011
	11111100000000110000000011111111 11111100000000110000000011111111 11110000000011110000000011111111 11110000000011110000000011111111 11110000000011000000000011111111 11110000000011000000000011111111 11110000000000000000001111111111 11110000000000000000111111111111 11110000000011111111111111111111 11110000001111111111111111111111 11110000000011111111111111111111 11110000000000000000001111111111 11110000000000000000001111111111 11110000000011000000001111111111 11111100000011000000001111111111 11111100000000000000001111111111 11111111000000000000111111111111 11111111110000000011111111111111 11111111000000000011111111111111 11110000000011110000000011111111 11110000001111111100000011111111 11110000001111111111000011111111	1111111110111111 1100001110001111 1110000100111111 1111000100011111 1110001100001111 1100001100001111 1100001000011111 1100001111111111 1100001111111111 1100001000011111 1110001000011111 1111000000111111 1111000001111111 1100011111001111

V. CONCLUSION

This paper exhibits feature extraction techniques for the symbols of Gujarati and Amharic scripts. Technique takes the scanned images of Amharic numerals as an input and produces wavelet compressed binrized editable file. As shown in the figure of section 3, Low-Low (Approximation) coefficients captures the core information of the original image with the size of around 32 x32 pixels. These coefficients are made up of just 256 (16 x 16) real numbers and hence compresses the original image to 75% or 1/4th in size. The last column of the table shown in above section shows the binarized coefficients with the threshold 0.90 i.e. if the coefficients are more than and equal to 0.90, replaced by 1 otherwise 0.These features can be used for the recognition purpose with neural network or other techniques [1, 4, 6].

REFERENCES

[1] Simon Haykin: Neural Networks A Comprehensive Foundation. 2nd edn. Prentice Hall, International, New Jersey (1999).

[2] Brijesh K. V.: Handwritten Hindi Character Recognition using Multilayer Perceptron and Radial Basis Function Neural Networks. IEEE International Conference, vol. 4. (1995) 2111-2115.

[3] Michael W. Frazier: An Introduction to Wavelets Through Linear Algebra. Springer- Verlag, New York (1999).

[4] Jacek M. Zurada: Introduction to Artificial Neural Systems. PWS Publishing Company (1992).

[5] Basit Hussain, M. R. Kabuka: A Novel Feature Recognition Neural Network and its Application to Character Recognition. Portal, IEEE Computer Society Washinton, DC, USA (1994) 98-106.

[6] Chi-Wei Lee and Zen Chen: Handwritten Chinese Character Recognition Using Stroke Structural Sequence Code. Journal of Information Science and Engineering, Vol. 13 No.4, Taiwan (1997) 631-648.

[7] Sameer Antani, Lalitha Agnihotri: Gujarati Character Recognition. Document Analysis and Recognition, IEEE (1999) 418-421.

[8] D. King, W. B. Lyons, C. Flanagan, E. Lewis: Signal processing technique utilising fourier transform methods and Artificial Neural Network pattern recognition for interpreting complex data from a multipoint optical fibre sensor system. ACM International Conference Proceeding Series, Trinity College Dublin (2004) 1-6.

[9] S. Rama Mohan, Archit Yajnik: "Gujarati Numeral Recognition Using Wavelets and Neural Network" Indian International Conference on Artificial Intelligence 2005, pp. 397-406.

[10] Archit Yajnik, S. Rama Mohan, "Identification of Gujarati characters using wavelets and neural networks" Artificial Intelligence and Soft Computing 2006, ACTA Press, pp. 150- 155.

[11] Dholakia J, Negi A, S. Rama Mohan, "Zone Identification in the Printed Gujarati Text", Proc. of 8th ICDAR, IEEE, 2005, pp. 272-276.