Amharic Handwritten Character Recognition Using Various Feature Extraction Techniques and Support Vector Machine

[1]Betselot Yewulu Reta, [2] Dhara Rana, [3] Gayatri Viral Bhalerao,[4]Melaku Eneayehu [5] Foram Soni
[1]M. Tech Student, [2]Lecturer, [3]Lecturer, [4]M. Tech Student, [5]M. Tech Student
[1]Department of Computer Science and Engineering,
[1]Parul Institute of Engineering and Technology, Vadodara, India

Abstract: Handwritten character recognition is an area of pattern recognition. It is the most challenge problem in pattern recognition. This paper presents the techniques for solving the challenges and problems of Amharic handwritten character recognition. In Amharic character, there is visual similarity of shape, and Amharic characters are large in number. In this paper, combined feature extraction techniques, such as zoning based and histogram of oriented gradients feature extraction techniques are used. Multiclass support vector machine used as a classifier. We trained and tested our approach on Amharic handwritten character data set and Chars74K benchmark numeric data set. we validated our model using 10-fold cross-validation

Index Terms **-Amharic handwritten character recognition, cross-validation, Error correcting output code, histogram of oriented gradients, support vector machine, optical character recognition, zoning.**

I. INTRODUCTION

In the development of technology soft copy documents converted to hardcopy. Similarly, hardcopy documents are also converted to machine readable format with the help of Optical character recognition(OCR). Optical character recognition is an area of Artificial intelligence and pattern recognition. It has different applications in banking, office, library system and so on. OCR also used in organization for data entry. It reduces cost of storage, and cost of accessing documents. By Converting hardcopy to softcopy using OCR method can decrease cost of accessing documents and storage. OCR can be classified into two categories namely handwritten character recognition and printed character recognition. In handwritten character recognition, inputs are handwritten characters. whereas in printed character recognition, inputs are machine printed characters. Handwritten character recognition also classified into two categories based on the type and acquirement of text namely online and offline character recognition. Input to online handwritten character recognition is sequences of strokes. Characters are distinguished by the time of writing and the order of strokes using specialized pen on digitized device. This specialized device controls the order of strokes. In offline handwritten character recognition, handwritten characters are available in the form of image. Offline handwritten character recognition is more challenging than online handwritten character recognition. Since different person uses different style, shape and orientation. Online handwritten character recognition identified by pen tip traces from pen-down to pen-up positions. In online handwritten character recognition, there is a special sensor which controls pen tip movements. For feature extraction pen pressure, velocity or change of writing direction are very important. Our focus is on handwritten character recognition for Amharic characters.

However, OCR system is strange for Amharic language. Amharic language is one of most spoken languages in an Afro-Asiatic language family, it belongs to under semantic language group. It is also the most spoken language next to Arabic in semantic language group. It is the official and working language of Ethiopia. Amharic language uses Ethiopic script as a writing style. Ethiopic script used by Amharic language are depicted in figure 1.1 which is the basic character used by Amharic language. As shown in the figure 1.1, the first column represents the base character, the rest six column represent their derived vocal sounds of base characters. Handwritten character recognition has been done in different language such as Hindi [1], Arabic [2], Bangla [3], Chinese [4] and Malayalam [5], and also some works has been done in Amharic language [7,8,9]. However, Amharic characters are large in number and there is visual similarity of shape with minor difference, further works need to be considered.

For Amharic handwritten character recognition, we proposed support vector machine as a classifier. SVM is a supervised machine learning algorithm used to separate the data points by maximizing the margin of hyperplane. Support vector machine used in different application area. Specially, in pattern recognition, support vector machine has been produced good result. It is robust to outliers. Some of the application are human detection [20], optical character recognition [1,14,15], baggage detection [21] and so on. In current trend, combined feature extraction approach produced good result [10,23]. For feature extraction, we applied combining feature extraction techniques. Histogram oriented gradient is one of the feature extraction techniques used in different application such as human detection [20], baggage detection [21], face recognition, character recognition [1] and so on. And also zoning feature extraction techniques is one of the powerful feature extraction technique and used in commercial OCR system [9,11]. We performed Amharic handwritten character recognition in each feature extraction technique. However, Amharic characters are similar with each other with

minor differences. So, we combining zoning based and histogram of oriented gradients feature extraction techniques. We trained and tested our approach on Amharic handwritten character data set and Chars74K benchmark numeric data set.

The paper is organized as follows: section 2 describes the challenges of Amharic handwritten character recognition. Section 3 describes review of papers. Section 4 describes data set preparation. Section 5 describes the general overview of the proposed system. Experiment and discussion details are described in section 6. Finally, section 7 concludes the paper.

		1st	2nd	3rd	4th	5th	6th	7th
		e/ä	u	i	a	ē	ə	o
1	h	ሀ	ሁ	ሂ	ሃ	ሄ	ህ	ሆ
2	l	ለ	ሉ	ሊ	ላ	ሌ	ል	ሎ
3	h	ሐ	ሑ	ሒ	ሓ	ሔ	ሕ	ሖ
4	m	መ	ሙ	ሚ	ማ	ሜ	ም	ሞ
5	s	ሠ	ሡ	ሢ	ሣ	ሤ	ሥ	ሦ
6	r	ረ	ሩ	ሪ	ራ	ሬ	ር	ሮ
7	s	ሰ	ሱ	ሲ	ሳ	ሴ	ስ	ሶ
8	sh	ሸ	ሹ	ሺ	ሻ	ሼ	ሽ	ሾ
9	q	ቀ	ቁ	ቂ	ቃ	ቄ	ቅ	ቆ
10	b	በ	ቡ	ቢ	ባ	ቤ	ብ	ቦ
11	v	ቨ	ቩ	ቪ	ቫ	ቬ	ቭ	ቮ
12	t	ተ	ቱ	ቲ	ታ	ቴ	ት	ቶ
13	ch	ቸ	ቹ	ቺ	ቻ	ቼ	ች	ቾ
14	h	ኀ	ኁ	ኂ	ኃ	ኄ	ኅ	ኆ
15	n	ነ	ኑ	ኒ	ና	ኔ	ን	ኖ
16	gn	ኘ	ኙ	ኚ	ኛ	ኜ	ኝ	ኞ
17		አ	ኡ	ኢ	ኣ	ኤ	እ	ኦ
18	k	ከ	ኩ	ኪ	ካ	ኬ	ክ	ኮ
19	h	ኸ	ኹ	ኺ	ኻ	ኼ	ኽ	ኾ
20	w	ወ	ዉ	ዊ	ዋ	ዌ	ው	ዎ
21		ዐ	ዑ	ዒ	ዓ	ዔ	ዕ	ዖ
22	z	ዘ	ዙ	ዚ	ዛ	ዜ	ዝ	ዞ
23	zh	ዠ	ዡ	ዢ	ዣ	ዤ	ዥ	ዦ
24	y	የ	ዩ	ዪ	ያ	ዬ	ይ	ዮ
25	d	ደ	ዱ	ዲ	ዳ	ዴ	ድ	ዶ
26	j	ጀ	ጁ	ጂ	ጃ	ጄ	ጅ	ጆ
27	g	ገ	ጉ	ጊ	ጋ	ጌ	ግ	ጎ
28	th	ጠ	ጡ	ጢ	ጣ	ጤ	ጥ	ጦ
29	ch	ጨ	ጩ	ጪ	ጫ	ጬ	ጭ	ጮ
30	ph	ጰ	ጱ	ጲ	ጳ	ጴ	ጵ	ጶ
31	ts	ጸ	ጹ	ጺ	ጻ	ጼ	ጽ	ጾ
32	ts	ፀ	ፁ	ፂ	ፃ	ፄ	ፅ	ፆ
33	f	ፈ	ፉ	ፊ	ፋ	ፌ	ፍ	ፎ
34	p	ፐ	ፑ	ፒ	ፓ	ፔ	ፕ	ፖ

Figure 1.1: Ethiopic script (Amharic characters)

II. CHALLENGES IN AMHARIC HANDWRITTEN CHARACTER RECOGNITION

Style variation, shape and orientation are not the only challenges in Amharic handwritten character recognition, but also similarity shape of characters and large number of characters are the critical issues in order to build an efficient OCR system. In general, the following are challenges or critical issues for Amharic handwritten character recognition:

1. There is visual similarity of characters. Amharic characters are very similar in shape with minor change (i.e. with the presences of special appendage of line, loop and dot below, above, left and right of the character). For example, character ሸ, ሹ, ሺ, ሻ, ሼ, ሽ, ሾ are very similar with each other. character ቀ, ቁ, ቂ ቃ, ቄ, ቅ, ቆ are very similar with each other. Character በ, ቡ, ቢ, ባ, ቤ, ብ, ቦ are very similar to each other. Similarly, all other Amharic characters are very similar with each other with minor difference.
2. Amharic characters are very large in number (totally 238 basic characters, 7x34=238 excluding special symbols and numbers). Computationally, it is difficult to train and test these large number of classes
3. There is no well-organized data set available online for Amharic characters.

III. RELATED WORK

In Amharic character, some works have been done in both real printed character and handwritten character. Because of the similarity of character in shape with one another, Amharic character recognition is still challenging and hot research area.

In paper [6], authors used two statistical algorithms to distinguish Amharic characters. The first approach is comparing Amharic characters with a series of templates. The second approach is creating signature from characters and compare with set of templates. In both algorithm, the characters are preprocessed (i.e. normalizing character size and adjusting orientation). The temple comparison techniques produce good result for very clear text.

In paper [7] proposed handwritten character recognition. It is based on structural features of primitive strokes, and the type of characters categorized based on their silent primitives. The structural and syntactic model grasp the orientation, relative length, structure and spatial position of primitive strokes by using directional field tensor. In this method, the authors build a primitive tree to hand the relationship, and the tree traversed to generate sequences of primitives. The sequence of strokes is compared against a knowledge base. The knowledge base stores sequences of primitive strokes and their connections of Ethiopic script.

In paper [8], authors used multiclass DAGSVM for recognizing real printed Amharic character (Ethiopic script). The authors performed feature extraction from entire image by concatenating all rows to create a single contiguous vector. The extracted feature consists of 0s for representing background and 1s for representing foreground pixels of the character image. For dimensional reduction of features, Principal Component Analysis(PCA) followed by Linear Discriminant Analysis(LDA) have been used. Finally, Multiclass DAGSVM applied to the given feature.

IV. DATA SET PREPARATION

Data set preparation is one of the most important task in OCR. We collected Amharic handwritten characters data set(AHCD) from three groups people by preparing A4 grid paper. In first group, we collected handwritten Amharic characters from 30 postgraduate students. In second group, we collected Amharic handwritten characters from 30 undergraduate students. Similarly, in third group, we collected Amharic handwritten characters from 30 native and non-native speakers of Amharic language with different background and age. The data set consists of 90-character images per class. We used Epson scanner to covert hardcopy to soft copy with 300dpi, and we used Gimp software to extract individual character. Finally, basic preprocessing techniques applied, such as gray scaling, binarization, thinning, noise removal, cropping, dilation and filling and normalization on collected isolated handwritten character data set. Below figure 4.1 is an example of the sample data set for the first fourteen Amharic handwritten characters. For additional experiment on our approach, we used Chars74K data set [22]. It is available freely online. It consists of Latin script and Hindu-Arabic numerals. Numeric character images and Amharic handwritten characters are cropped in to 64x128 window (64 pixels wide by 128 tall).

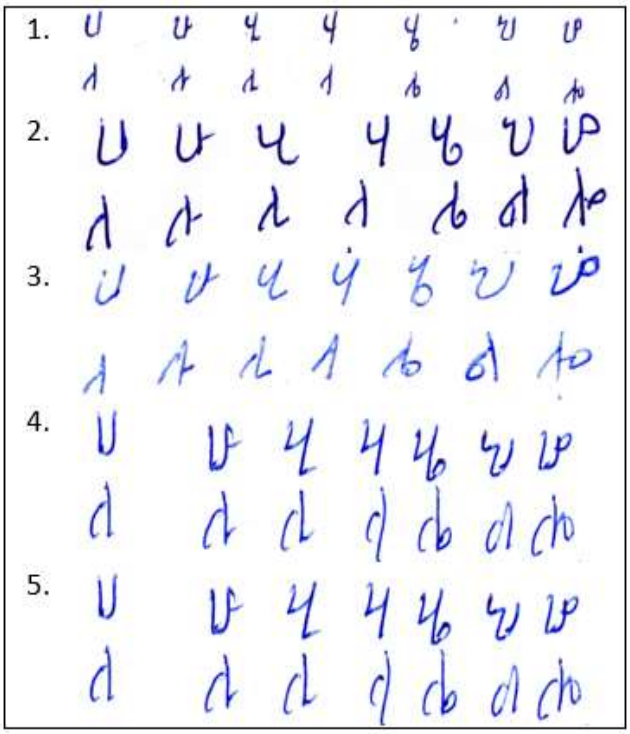


Figure 4.1: Sample dataset for the first fourteen Amharic handwritten characters (for five persons)

V. OVERVIEW OF THE PROPOSED SYSTEM

The proposed system for Amharic handwritten character recognition consists of four main steps. In the first step, character images are collected, and used as input to preprocessing steps. In this step, basic preprocessing techniques are applied. These prepressed characters used as input for feature extraction steps. In feature extraction, zoning based and HOG feature extraction techniques are

applied. Finally, Support vector machine used as a classifier. Generally, the proposed system described in figure 5.1 using block diagram.

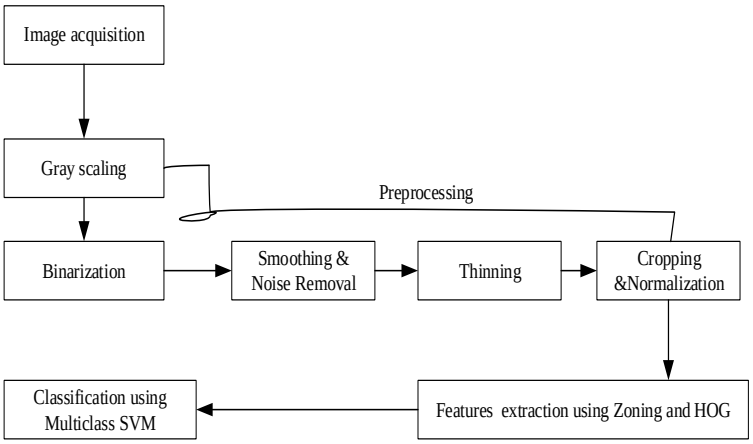


Figure 5.1: Architecture of the proposed Amharic handwritten character recognition system

5.1 Image Acquisition

Image acquisition is the processes of collecting images (in our case image of characters) using digital camera and scanner. In this phase, we collected handwritten Amharic characters using scanner. We used Epson scanner with 300dpi

5.2 Preprocessing

Preprocessing is the basic step of character recognition since different noise exist due to poor quality of scanner and paper. The aim of preprocessing is to remove noise and to make the subsequent steps simple. It includes cropping, normalization, gray scaling, binarization, smoothing and noise removal, thinning, dilation and filling and so on. For binarization we applied Otsu's global threshold, for thinning Zhang-Suen parallel thinning algorithm [24], and for smoothing and noise removal, we applied median filtering.

5.3 Feature Extraction

Feature extraction is the basic step in optical character recognition to achieve good performance of recognition [11]. The basic features of character extracted using different feature extraction technique and stored in feature vector. The extracted feature vector used as input to the classification algorithm. In general, properly extracted feature plays significant role to leverage the performance of classification algorithm [10].

5.3.1 Zoning Based

Zoning based feature extraction is widely used techniques in handwritten character recognition. In zoning, the character image divided into a number of zones and features are extracted from each zone. Character images divided into overlapping and non-over lapping regions [12]. In each zone: Average distance from the character centroid to each pixel present in the zone is to be computed, similarly zone centroid is computed and average distance from the zone centroid to each pixel present in the zone is computed, pixel density calculated in each zone and area /number of pixels also computed in each zone. In each zone, structural feature of the character extracted, such as horizontal line, vertical line, right diagonal and left diagonal [12]. We divide the image into 16 regions with size of 32x16 pixels. Rather than extracting features from the entire image, we extract features from 16 regions. Thus, the detail features are extracted from each region. Zonal feature extraction is produce good result for characters having similar shapes. Since when we dividing the character image into zones, the different parts of the character fail in different regions. Let us consider the two Amharic characters as shown in figure 5.1. The two characters are the same except the appendage of horizontal line at the bottom. In figure 5.1(a), the horizontal line or appendage of line fail in the region from (4,3) to (4,4). Which means the difference of the two character is the presence of horizontal line at the region of (4,3) and (4,4) (in the first character image, but there is no horizontal line in the second figure 5.1(b) in the region of (4,3) and (4,4). So, in each region, the different features are extracted to differentiate similar characters. As we stated, Amharic characters are similar in shape with each other with slight difference. Some zones are empty, so these zones have zero values in feature vector.

5.3.1.1 Universe of Discourse

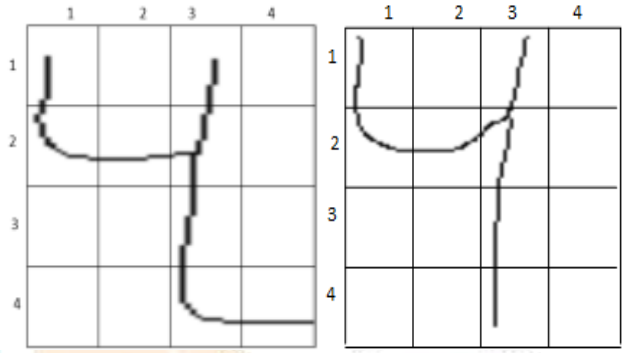
In order to extract features using zoning-based technique, we defined a short matrix, which fits the entire character skeleton. Every character is independent of image size.

5.3.1.2 Starters, Intersection and Minor Starters

In order to extract different line segment, in specified zone, the entire character skeleton in that zone should be transverse. So, to do this task, certain pixels defined as starters, intersections and minor starters. More expiation is available [9]. After extracting line segments, it is categorized under horizontal line, vertical line, left diagonal line, right diagonal lines and curve. Directional vector is used to determine each line type. After identifying the line type the next step is to build feature vector. In each zone, different line segments extracted and stored in feature vector. These features extracted from each zone are: number of horizontal lines, number of vertical lines, number of curve lines, number of diagonal lines, number of right diagonal lines, number of left diagonal lines, normalized length of all horizontal lines, normalized length of all curve lines, normalized length of vertical lines, normalized length of left diagonal lines and normalized length of right diagonal lines.

The number of each line type is normalized with,

value = 1-((numberOfLines/10)-2
Normalized length of each line type is obtained by,
length=(Total Pixels in that line type)/(total zone pixels).
In addition to, structural feature we also used other features such as Euler number, area and eccentricity.
Euler number: Is the difference of the number of object in the image and the number of holes in the image. Some of Amharic characters have holes or loops, so in order to distinguish characters having holes or loops and do not have holes or loops. For example the Euler number of characters, such as ሄ, ሆ, ሌ, ሎ, ሜ, and ሤ is 0(1-1=0) and ሁ, ሂ, ሉ, ሊ, ሚ,ሢ becomes 1(1-0=1). This technique has been applied for the rest of Amharic characters.
Area: Is the number of pixels in a particular region. We calculated area both in each zone and entire image.
Eccentricity: It can be expressed in the ratio of foci of ellipse and its major axis.



a. b.

Figure 5.2: Zoning-based representation of character hi(ሂ) (a) and ha(ሃ) (b)

5.3.2 Histogram of Oriented Gradients

Histogram of oriented gradients used as a feature descriptor by computing the occurrence of magnitude and gradient orientation of localized part of an image [10]. One of the best feature HOG are easy to express the rough structure of the object and invariant in geometry, contrast and illumination changes. HOG feature descriptor used a specified matrix. we used sobel operator, which slides over the entire character image in order to compute gradient orientation and magnitude. Generally, HOG features are calculated as follows: first the character image partitioned in to a number of cells of size 8x8 pixels. Gradient magnitude and gradient orientation computed for every pixel in each cell. Each gradient magnitude discretized into histogram bins based on gradient orientation. In order to alleviate contrast and illumination problems, block normalization is applied. Block normalization is performed on overlapping and non-over lapping blocks. Finally, these normalized blocks are used as HOG descriptor.

5.3.3 Combined Features

There are different feature fusion techniques, such as serial feature fusion, parallel feature fusion and feature fusion based on Canonical Correlation Analysis (CCA) and Discriminant Correlation Analysis (DCA). Serial feature fusion is simple combining two set of features vectors into one feature vectors. Parallel feature fusion is used to combine two features into a complex feature vector. Whereas feature fusion based on Canonical Correlation Analysis (CCA) and Discriminant Correlation Analysis(DCA) is used the relationship between two set of features in order to get two set of transformations by maximizing the correlation of transformed feature across the two feature sets and by discriminating within each feature sets [13]. we applied feature level fusion using Discriminant Correlation Analysis(DCA). DCA used in supervised learning by keeping the class structure.

5.4 Classification

After feature extraction of characters, the next step is classification. At this stage the decision has been made based on the feature extracted from characters. Classification methods are used to apply in pattern recognition to separate one class from the rest of classes. Classification method can be statistical approach [1,2], template matching [6], syntactic and artificial neural network [4,5].

5.4.1 Support Vector Machine

First introduced by Boser, Guyon and Vapnik in 1992. Support vector machine is a supervised machine learning algorithm used to solve linear and nonlinear classification problem [14]. SVM is based on the structural risk minimization, it is better than empirical risk minimization used in neural network [14]. The aim of SVM is to find a parameter setting that decrease the risk famulated by,

$$R_{emp}(\alpha) = \frac{1}{2l}\sum_{i=1}^{l} | y_i - f(x_i, \alpha) | \qquad (5.1)$$

Where yi represent the output, xi is represent the input and α is parameters.

Support vector machine used for classification and regression problems. It plots the data as a point in multidimensional space. After plotting the data items on n-dimensional space, it classifies the data by drawing a hyperplane that differentiate the classes. In SVM, the hyperplane is drawn based on the maximum margin. The best characteristics of support vector machine is robust to outlier. SVM find a global minimum [15]. There are two cases in which SVM is used. The first case is separable. Label data such as $\{x_i,y_i\}$, i=1,2 ,…l, $y_i \in \{-1,1\}, x_i \in R^d$. There is hyperplane that separate the positive class from negative classes. The point x satisfies w.x +b=0. Where w

is a weight vector, |b|/||w|| is the distance from the hyperplane to the origin, and also ||w|| is Euclidian norm of w. In linear separable case, SVM selects a hyperplane with largest margin. The equations are:

$$x_i.w+b \geq +1 \quad for \quad y_i = +1 \tag{5.2}$$

$$x_i.w+b \leq -1 \quad for \quad y_i = -1 \tag{5.3}$$

The combination of the two equation is

$$x_i.w+b-1 \geq 0 \forall_i \tag{5.4}$$

Consider (5.4), the points lie on hyperplane h1: xi.w+b=1 with normal w, perpendicular distance is |1-b|/||w|| and in (5.5), the points lie on hyperplane h2:xi.w+b=-1 with norm w and the distance from the origin |-1-b|/|b||. In general, we find hyperplanes that gives largest margin by maximizing $|| w ||^2$, based on (5.4),

$$w = \sum_i \alpha_i y_i x_i \tag{5.5}$$

Where α_i is Lagrange multiplier for every training points. All points, $\alpha_i > 0$ are support vectors. The rest training points have $\alpha_i = 0$.

The second case, when the data is non-separable. The above equation not produce feasible result. Non-negative slack variable is introduced, ξ_i $i = 1, 2, \ldots, l$, so the linear separable equation changed to the following equation:

$$y_i(x_i.w+b) \geq 1-\xi_i \ \ for\ i = 1,2,..l \tag{5.6}$$

The non-separable problem becomes

$$\text{Min}\ \frac{1}{2} || w ||^2 + C\sum_i \xi i \tag{5.7}$$

$$\text{Subject to } y_i(x_i.w+b) \geq 1-\xi_i \ and\ \xi_i \geq 0 \ \ i = 1,2,..l$$

C is the parameter to be chosen by the user. In this case w is given by (5.8),

$$w = \sum_i \alpha_i y_i x_i \tag{5.8}$$

In this case, the difference between (5.5) and (5.8) is that α_i have an upper bound of C.

The data is mapped into infinite dimensional Euclidian space H. using Φ .

$$\Phi : R^d \rightarrow H \tag{5.9}$$

The training algorithm depends on data using dot products. The kernel function k can be used in support vector machine for transforming low dimension in to higher dimensions, $k(x_i, x_j) = \Phi(x_i).\Phi(x_j)$ .where k is representing kernel function. There are tuning parameters in SVM such as kernel, regularization, gamma and margin. The most common Kernel functions are:

- Linear kernel: $k(x_i, x_j) = x_i \times x_j$
- Polynomial: $k(x_i, x_j) = [(x_i \times x_j)+1]^d$
- Sigmoid: $k(x_i, x_j) = \tanh(\beta_0 x_i x_j + \beta_1)$
- ERBF: $k(x_i, x_j) = \exp(|| x_i - x_j ||)/2\sigma^2$

d , β_0, β_1 and σ are parameters to be determined empirically. The test phase of an SVM is used by computing the sign of (5.10),

$$f(x) = \sum_{i=1}^{Ns} \alpha_i y_i \Phi(s_i).\Phi(x) + b = \sum_{i=1}^{Ns} \alpha_i y_i K(s_i, x) + b \tag{5.10}$$

Where Si are support vectors.

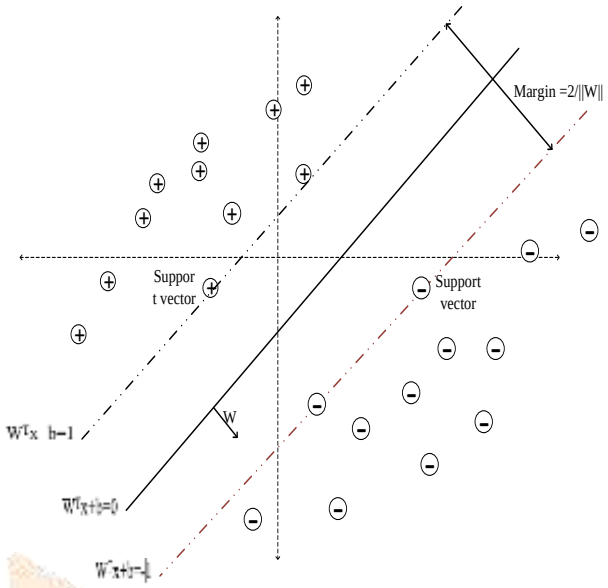


Figure 5.3: SVM for binary classes

Binary classification problems have been studied in different literatures. And also, multiclass classification has been introduced to solve multiclass classification problem. There are two approaches for multiclass classification problem, such as direct multiclass design and (indirect) decomposition design. Decision tree and neural network are examples of direct approach. whereas multiclass SVM such as one versus one, one versus all and ECOC are example of indirect approach. Multiclass classifier works by combining multiple binary classifiers. In multiclass classification, the most important techniques have been introduced, such as one versus all(OVA) [15], one versus one(OVO) [16] and error correcting output code(ECOC) [16, 17]. Suppose N classes, OVA train one binary classifier per class, totally N binary classifier created. In OVA, if one class c is labeled as positive the other classes are labeled as negative. Whereas OVO is pairwise classifier. Suppose there is problem with c different classes, so, in OVO, c(c-1)/2 classifiers are trained for separating observation of one class from the rest of observation of classes. Unknown input can be classified based on the maximum voting, each classifier votes for single class.

Error correcting output code (ECOC) is one of the multiclass classification techniques which produce the successful result. ECOC is a framework for multiclass classification, used to increase the performance of the base class. ECOC has three components: coding, binary classifier and decoding. In coding techniques, coding matrix is created for the given problem [16]. The coding matrix contain column and row. Each row depicts the codeword for each class. And the column depicts the classifiers. Coding is used where the binary problem has been dealt and designed. Coding is divided into two categories, such as binary coding and ternary coding based on the membership to binary or ternary ECOC [14]. The most common coding design strategy are One versus all and one versus one. In one versus all strategy, each binary classifier is trained to distinguish one class from the rest of the classes. For a given class N, one versus one has N bit codewords [14]. One versus one strategy considers all possible praise of classes [14]. Its codeword length could be N(N-1)/2. Binary classifier consists of group of independent binary classifiers that are trained based on different categories of the original data and each column of the coding matrix. Whereas decoding is the final classification produced based on the result of binary classifiers. It is the problem of determining the distance between the test codewords and codewords of the classes. The most common decoding techniques are hamming decoding, Euclidean decoding, lose based decoding and so on.

In hamming decoding, assume the hamming distance between class codewords is m, (m-1)/2 errors in individual base classifier result can be adjusted, because the minimum distance of the codeword will be the correct one [17]. In ECOC framework, create a codeword for each class, the codeword is organized as a matrix. Where the matrix M Є {-1, 0, 1} N*n of the ternary case, n is the code length [17]. The row of the code matrix depicts the class and the column represent binary classifier. -1 represent the class considered as negative by the classifier, 1 represent the class considered as positive and 0 represent the class is not considered by the classifier. In the decoding part, the hamming distance is given by,

$$d(x, y^i) = \sum_{j=1}^{n} | x_j - y_j^i | / 2 \qquad (5.11)$$

Where x is the value input vector codeword and y is the value of base class codeword.

The decoding techniques is based on the error correcting principles under the assumption that the learning task can be designed as a communication problem. Which means the class information is transmitted over the channel, and the two symbols could be found at each position of the sequences [14].

Euclidian decoding is s one of decoding technique defined by,

$$d(x, y^i) = \sqrt{\sum_{j=1}^{n} (x_j - y_j^i)^2} \qquad (5.12)$$

Whereas, in loss-based decoding techniques, each binary classifier result a margin score based on two requirements [18]. First, the score should be positive if the example is categorized as positive and negative if the example is categorized as negative. Second, the magnitude of the score measure of confidence in the prediction. Let f(x, b) be the margin score predicted by dichotomizer corresponding to the column b of the given code matrix for example x. For each row c of the matrix M for each example x, the distance can be computed between f and M (c, b) as,

$$dL(x,c) = \sum_{b=1}^{l} L(M(c,b)\, f(x,b)) \qquad (5.13)$$

Where L is loss function and M(c, b)=0,1or -1. Each example x is labeled with the label c* in which dLis minimized.

VI. EXPERIMENT AND DISCUSSION

For the experiment, we used MATLAB programming language. The implementation is done on 8GB RAM, core i7 processor and 2TB hard disc dell laptop. Both numeric data set and Amharic handwritten data set characters cropped into 64x128 window (64 pixels wide by 128 pixels tall). Features are extracted in this cropped character image. There are different parameters of multiclass SVM and HOG descriptor in which we performed. HOG parameters such as cell size is 8x8 pixels, block size is 16x16 pixels, orientation of 9 bins is 0^0-180^0, for calculating gradient we used sobel operator, and we normalized block histogram with L2-Norm. And also, in zone-based feature extraction, the entire image is divided into four regions of size 32 by 16 pixels and features are extracted in each region. Similarly, in multiclass SVM, we applied linear kernel function and one versus one coding matrix. The model is evaluated using 10-fold cross-validation to eradicate overfitting problem. In HOG, when the cell size is 2x2 and 4x4 pixels, the accuracy decreased and the dimension of feature increased in comparison with cell size of 8x8 pixels. 8x8 pixels cell size produced good result than 2x2 and 4x4 pixels of cell size for our approach. In order to eradicate illumination, contrast and shadow, we normalized the histogram on overlapping and non-overlapping blocks of size 16x16 pixels using L2-norm. Polynomial and gaussian kernel are also used by changing order and gamma, respectively, but the accuracy is very low compared to linear kernel. Linear kernel function produced good result in Amharic handwritten character recognition. In designing code matrix for ECOC multiclass framework, one versus all better than one versus one encoding. One versus one computationally expensive than one versus all. From the experimentation, we observed that using HOG as feature descriptor and SVM with linear kernel as classifier produced an accuracy of 80.9%. Using zoning and geometrical feature, we achieved an accuracy of 69.8%. In both approaches, the accuracy is not satisfactory, since in Amharic character, there is visual similarity of shape. The fusion of HOG and zoning-based produce good result in recognition of Amharic handwritten characters. In fusion approach, the accuracy is 90.8%. It shows that, combined approach better than single approach. To combine the two features, we applied feature level fusion based on Discriminant Correlation Analysis(DCA) [13]. Amharic characters are large in number so it is difficult to summarize in graphically and using confusing matrix. we created 238 by 238-dimension confusion matrix using MATLAB. We obtained the overall accuracy from confusion matrix for all approach and we summarized using table 1 shown below. Similarly, we applied our approach on Chars74K benchmark numeric data set, the result is shown in table 2. From the experiment on Chars74K benchmark numeric data set, we observed that there is incorrect classification between, 0 and 9 and 4 and 9, and for other numeric characters described using confusion matrix figure 6.1 shown below. The overall accuracy on Chars74K benchmark numeric data set is 92.0%.

6.1 Cross-validation

Cross-validation is one of the model evaluation technique, used to evaluate how the learner will generalize or predicate unseen or new data. It is a rotation estimation method [19].

In cross validation data set is randomly divided into k mutually exclusive folds d_1, d_2, d_k of approximately equal size. Thus, the model trained and tested k times, each time tϵ {1, 2,…, k}, it is trained on d/ d_t tested on d_t. the accuracy can be calculated by dividing the overall number of correct classification of the model over the number of instance of the dataset. Cross-validation estimate is a random number which depends on the division of data sets into folds. There are different cross-validation techniques, such as holdout, k-fold cross-validation and leave-one-out cross validation [19].

Table 6.1: Accuracy of recognition using different feature extraction techniques on AHCD data set

Sr. No.	*Feature extraction techniques*	Classification algorithm	*Coding matrix*	*Overall Accuracy*
1	HOG feature	SVM with liner kernel	One versus one	80.9%
2	Zoning and geometrical	SVM with liner kernel	One versus one	69.8%
3	Fusion of HOG and Zoning	SVM with liner kernel	One versus one	88.5%
4	HOG feature	SVM with liner kernel	One versus all	83.7%
5	Zoning and geometrical features	SVM with liner kernel	One versus all	73.0%
6	**Fusion of HOG and Zoning**	**SVM with liner kernel**	**One versus all**	**90.8%**

Table 6.2: Accuracy of recognition using different feature extraction techniques on Chars74K benchmark numeric data set

Sr. No.	*Feature extraction techniques*	Classification algorithm	*Coding matrix*	*Overall Accuracy*
1	HOG feature	SVM with liner kernel	One versus one	82.6%
2	Zoning and geometrical	SVM with liner kernel	One versus one	72.4%
3	Fusion of HOG and Zoning	SVM with liner kernel	One versus one	90.0%
4	HOG feature	SVM with liner kernel	One versus all	85.3%
5	Zoning and geometrical	SVM with liner kernel	One versus all	75.9%
6	**Fusion of HOG and Zoning**	**SVM with liner kernel**	**One versus all**	**92. 0%**

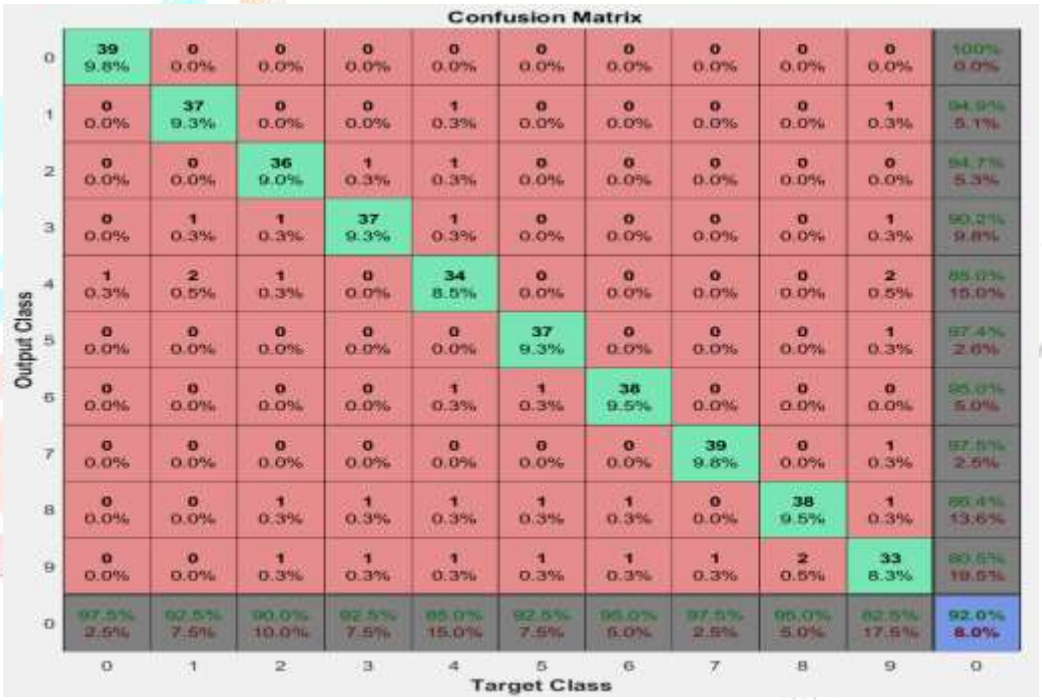


Figure 6.1: Confusion matrix using fusion of HOG and zoning features and SVM on Chars74K benchmark numeric data set.

VII. CONCLUSION

Handwritten character recognition is on the interesting and challenging area in pattern recognition. There are some challenges in Amharic characters. Amharic characters are similar with minor difference, such as the presence of small lines below, above, to the right, to the left of the character. In addition to visual similarity of characters, Amharic characters are large in number (238 basic characters excluding infrequently used characters, special symbols and numbers). In this paper, the various feature extraction techniques such as HOG, zoning and structural and geometrical features are applied. In zoning based, character images are partitioned into 16 zones of size 32x16 pixels, in each zone various structural features are extracted such as, horizontal line, vertical line, diagonal line, curve line and area. Whereas geometrical features such as Euler number, area, eccentricity, extent and eccentricity are extracted from the entire character image. Similarly, In HOG feature extraction techniques, gradient magnitude and gradient orientation computed by dividing the character image in different cells. Histogram and block normalization applied in each cell. We trained in each feature and also, we trained by fusing of HOG and Zoning-based feature and multiclass SVM as classifier. The combined approach gave us good result in both Amharic handwritten dataset(AHCD) and Chars74K benchmark numeric dataset. The model is validated using 10-fold cross validation techniques. In future work, we will focus on distinguishing visual similarity of Amharic characters and increasing our data set. Furthermore, we suggest that, applying other machine learning algorithms such as artificial neural network or deep learning, the accuracy will be improved

VIII. ACKNOWLEDGMENT

The authors would like to thank department of computer science and engineering, Parul institute of engineering and technology, Parul university, their valuable information and support for this work.

REFERENCES

[1] A. Gaur and S. Yadav, "Handwritten Hindi character recognition using k-means clustering and SVM," 2015 4th International Symposium on Emerging Trends and Technologies in Libraries and Information Services, Noida, 2015, pp. 65-70.
[2] El Moubtahij, Halli and Satori, "Recognition of Off Line Arabic Handwriting Words Using Hidden Markova Model(HMM) Toolkit," 3th International Conference on Computer Graphics, Imaging and Visualization (CGiV), Beni Mellal, 2016 pp.167-171.
[3] M. M. R. Sazal, S. K. Biswas, M. F. Amin and K. Murase, "Bangla handwritten character recognition using deep belief network," 2013 International Conference on Electrical Information and Communication Technology (EICT), Khulna, 2014, pp. 1-5.
[4] S. Yang, F. Nian and T. Li, "A light and discriminative deep networks for off-line handwritten Chinese character recognition," 32nd Youth Academic Annual Conference of Chinese Association of Automation (YAC), Hefei, 2017, pp. 785-790.
[5] P. Naire, A. Jamese and C. Saravanan, "Malayalam handwritten character recognition using CNN," International Conference on Inventive Communication and Computational Technologies (ICICCT), Coimbatore, 2017, pp. 278-281.
[6] J. Cowell and F. Hussain, "Amharic character recognition using a fast signature-based algorithm," Proceedings on Seventh International Conference on Information Visualization, 2003. IV 2003., 2003, pp. 384-389.
[7] Y. Assabie and J. Bigun, "HMM Based Handwritten Amharic Word Recognition with Feature Concatenation," 10th International Conference on Document Analysis and Recognition, Barcelona, 2009, pp. 961-965.
[8] M. Meshesha and C. V. Jawahar, "Recognition of printed Amharic documents," Eighth International Conference on Document Analysis and Recognition (ICDAR'05), 2005, pp. 784-788.
[9] Dinesh Dileep, "A feature extraction technique based on character geometry for character recognition"
[10] Betselot Y. Reta and Sheetal M. Thakar," A Survey on Understanding Handwritten Character Recognition Feature Extraction Techniques," Parul University International Conference on Engineering & Technology(PiCET-2018): Smart Computing, Vadodara, India, February 2018.
[11] Ashlin Deepa and R.N, R.RajeswaraRao, "Feature Extraction Techniques for Recognition of malayalam Handwritten characters "International Journal of Advanced Trends in Computer Science and Engineering, Vol. 3, No.1, 2014,pp. 481– 485.
[12] Purna Vithlani and C.K.Kumbharana "Structural and Statistical Feature Extraction Methods for Character and Digit Recognition" Int. Journal of Computer Applications Volume 120, No.24, 2015,pp. 0975 – 8887.
[13] M. Haghighat, M. Abdel-Mottaleb and W. Alhalabi, "Discriminant correlation analysis for feature level fusion with application to multimodal biometrics," 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Shanghai, 2016, pp. 1866-1870
[14] R. M. J. S. Bautista, V. J. L. Navata, A. H. Ng, M. T. S. Santos, J. D. Albao and E. A. Roxas, "Recognition of handwritten alphanumeric characters using Projection Histogram and Support Vector Machine," International Conference on Humanoid, Nanotechnology, Information Technology,Communication and Control, Environment and Management (HNICEM), Cebu City, 2015.,pp. 1-6
[15] A. Mowlaei and K. Faez, "Recognition of isolated handwritten Persian/Arabic characters and numerals using support vector machines," IEEE XIII Workshop on Neural Networks for Signal Processing (IEEE Cat. No.03TH8718), 2003, pp. 547-554.
[16] M. a. Bagheri, G. A. Montazer and S. Escalera, "Error correcting output codes for multiclass classification: Application to two image vision problems," The 16th CSI International Symposium on Artificial Intelligence and Signal Processing (AISP 2012), Shiraz, Fars, 2012, pp. 508-513.
[17] Thomas G.Dietterich and Ghulum Bakiri" Solving Multiclass Learning Problems via Error-Correcting Output Codes," journal of Artificial Intelligence Research 2, USA , 1995, pp. 263-286.
[18] Allwein, E., R. Schapire, and Y. Singer. "Reducing multiclass to binary: unifying approach for margin classifiers." Journal of Machine Learning Research. Vol. 1, 2000, pp. 113–141.
[19] Ron Kohavi, "A study of cross-validation and bootstrap for accuracy estimation and model selection," IJCAI95 Proceedings of the 14th international joint conference on Artificial intelligence -vol.2, 1995, pp.1137-1143.
[20] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), San Diego, CA, USA, vol. 1, 2005, pp. 886-893.
[21] T. Khanam, K. Deb and K. H. Jo, "Baggage detection and classification using human body parameter & boosting technique," 10th International Conference on Human System Interactions (HSI), Ulsan, 2017, pp. 54-59.
[22] T. E. de Campos, B. R. Babu, and M. Varma, "Character recognition in natural images," in Proceedings of the International Conference on Computer Vision Theory and Applications, Lisbon, Portugal, February 2009.
[23] M. Yadav and R. Purwar, "Hindi handwritten character recognition using multiple classifiers," 2017 7th International Conference on Cloud Computing, Data Science & Engineering - Confluence, Noida, 2017, pp. 149-154.
[24] T. Y. Zhang and C. Y. Suen, "A fast parallel algorithm for thinning digital patterns," Communications of the ACM, vol. 27, 1984, pp.236-239