Amharic Handwritten Character Recognition using Machine Learning Approach

[1]Betselot Yewulu Reta, [2]Dhara Rana, [3]Gayatri Viral Bhalerao, [3]Melkye Wereta Tsigie
[1]M. Tech Student, [2]Lecturer, [3] Lecturer
[1] Department of Computer Science and Engineering,
[1] Parul Institute of Engineering and Technology, Vadodara, India

Abstract: Handwritten character recognition is one of the most challenging problem in the area of pattern recognition. Since different persons have different writing styles. And also, Amharic characters are large in number and some of the characters shape are similar with minor change. In this paper, the basic character recognition techniques performed, such as preprocessing, feature extraction and classification. After preprocessing, haar wavelet transform followed by histogram-oriented gradients feature extracted in each character image. Using haar wavelet transform the character image decomposed into four coefficients such as average, horizontal, vertical and diagonal. Then, HOG feature extracted from four coefficients. The final feature vector created as a single feature vector by concatenating the HOG features extracted from each coefficient. The features used by HOG are gradient orientation and magnitude. The dimension of feature is reduced using LDA. A special type of multiclass SVM in ECOC framework with one versus all design coding matrix is adopted. The algorithm is trained and tested on Amharic handwritten characters data set and Chars74K benchmark numeric data set. The proposed model is validated using 10-fold cross-validation technique. As a result, Multiclass SVM classification algorithm and haar wavelet followed by HOG feature extraction technique have been achieved good result in recognizing Amharic handwritten characters.

Index Terms **– Amharic handwritten character recognition; Error Correcting Output Code; Haar Wavelet transform; histogram of oriented gradients; linear discriminate analysis; optical character recognition; support vector machine;**

I. INTRODUCTION

In advancement of technology soft copy documents converted to hard copy using special device called printer and similarly handwritten scanned document and real printed scanned document can be converted in to machine readable ASCII format with the help of Optical Character Recognition (OCR). OCR is a technique used for digitizing scanned handwritten and printed text into ASCII format. OCR is one of the challenging research area in pattern recognition. OCR system can be classified into two categories: Printed Character Recognition(PCR) and Handwritten Character Recognition(HCR). The input to Printed character recognition is scanned real machine printed character images, whereas the input to handwritten character recognition is scanned handwritten character images. Handwritten character recognition also can be classified into two categories based on the type and acquirement of text: offline and online character recognition. In Online character recognition, characters are recognized at the time of writing. The input to online character recognition is sequences of strokes. The characters are identified by the time and the order of strokes using specialized pen on digitized device. Similarly, in offline character recognition, handwritten characters are scanned and available in the form of image. Offline character recognition is more challenging than online character recognition. In addition to the different writing style, shape and orientation of different person, handwritten characters may contain noise. The noise in handwritten characters can decrease the performance of recognition. OCR has different application: In Banking, OCR system used for data entry by scanning checks from a printed document and also handwritten document with minimum human intervention. In legal and library system, OCR is also used as data entry by scanning books and printed materials. As a result, the document is stored in compact storage, and accessed in easily from the huge library in electrical format. And also reading Books for blinded people and barcode reader are some of the applications of OCR. The application of OCR is countless. It reduces cost of storage, and cost of accessing documents. OCR is under the category of pattern recognition, artificial intelligence and computer vision.

However, OCR system is peculiar for Amharic language. Amharic language is one of the most spoken languages in an Afro-Asiatic language, which belongs to Semantic group. In semantic language family, Amharic is the most spoken language next to the Arabic language in the world and also the largest language in Horn of Africa[20]. Amharic language has its own writing system which is called Ethiopic script. Ethiopic script has 34 base characters with six orders, depicts the derived vocal sound of base characters [20]. Ethiopic script is illustrated in fig 1. The first column represents base characters and the rest represent their derived vocal sounds.

Handwritten character recognition has been done in Latin characters [23,24] and also in non- Latin characters such as Malayalam [6], Hindi [28], Arabic [2,4,9], Bangla [3,15,42], Chinese [5,21]. In Amharic language also, some works have been done in real printed character recognition [17] and handwritten word recognition [18,19,20]. For Amharic Handwritten character recognition, we proposed a machine learning approach. Till now, SVM technique has not been used in handwritten Amharic characters. In this study, SVM classifier for Amharic handwritten character recognition have proposed. As a result, the important features of SVM is studied. Currently, support vector machine (SVM) achieves better accuracy in pattern recognition like character recognition. Support vector machine is one of the supervised classification algorithms. It works by maximizing the margin that separate the positive class from negative class. It also transforms a low dimensional vector space to higher dimensional vector space using kernel functions. Support vector machine used in both binary class problem and multiclass problem. In binary class classification problem, SVM used to separate the two classes such as email spam detection, medical diagnosis, quality control in factories (i.e. to decides the specification

has or has not met), and others. Whereas multiclass classification problem is used to classify instances into three or more classes. As a result, SVM is used in multiclass classification problems. SVM achieved by decomposing the multiclass classification problems into binary class classification problems. Some of the approaches are one vs one, one versus all and Error Correcting Output Code (ECOC). ECOC with one versus all coding design is adopted. One versus all is still the optimal method of multiclass classification problems [27].

		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

Fig. 1. Ethiopic Script (Amharic characters)

1.1 **Challenges in Amharic Handwritten Character Recognition**

Handwritten character recognition for Amharic character is the most challenge problem because Amharic characters are similar in shape with minor change. In Ethiopic script (Amharic character), there are 34 base characters with six orders. The six orders represent the derived vocal sounds of the base characters. Within each base character and derived characters, there is similarity of shape with minor or slight change. And also, there is similarity of shape between/among base characters and their derived characters with minor difference.

II. RELATED WORK

In [17], they proposed two statistical algorithms to distinguish Amharic characters. The first approach is comparing Amharic characters with templates. The second approach is creating signature for characters and compare with set of templates. In both

algorithm, the characters are preprocessed. The authors conducted the experimental on both approaches. Thus, signature approach is 50 times faster than the original character and recognition process. The signature is derived from normalized characters.

In this paper [27], the authors employed multiple feature and neural network for Devanagari character. In this study, 8 structural features are extracted by partitioning the image into nine zones. Totally 72 structural features are extracted (8x9). By the same token, 9 global features are obtained, such as Euler number, convex area, filled area, solidity, perimeter, area, eccentricity, extent and orientation. Multilayer perceptron neural network is used recognition of characters. The network consists of 81 neurons in input layer, varied number of neurons in hidden layer and 10 output neurons in output layer for numerical characters. The accuracy achieved good result.

In this paper [28], the author employed artificial neural network for handwritten digit recognition system for south Indian language. In this study, ANN is used as classifier and HOG feature is used as a feature extractor of the character. Each image has size of 130 x 66 pixels. In firs step of HOG, Gradient values are computed by using one dimensional masks. The HOG feature works on an image size of 128x64 pixels. The system used detection window size of 8 x 8 pixels (i.e. cells). The cells are encircled into overlapping blocks with 50%. Each block composed of four cells. Each block touch half of the preceding block. As a result, seven horizontal and fifteen vertical blocks are formed. 64 gradient vectors are calculated from each cell. These vectors are placed on the histogram of nine bins. In the histogram, angle of gradients are depicted on x-axis and magnitudes are depicted on y-axis. The second step in HOG is block normalization. Block normalization are performed in each histogram of four cells. The histogram four cells in block are combined into a one vector with 36 values. These values are normalized by dividing it with magnitude. Finally, 105 blocks formed (7x15). The descriptor has 3,780 values. ANN is used finally to recognizing the character on 2500 sample images of size 130 x 66 pixels. The accuracy produced good result.

In Paper [30], the author proposed handwritten alphanumeric character recognition using projection histogram and support vector machine. The authors follow basic optical character recognition steps: preprocessing feature extraction, and classification. In preprocessing, basic preprocessing operations performed such as, gray scaling, noise removal, binarization and others. Similarly, in feature extraction, project histogram is counting pixel distribution in horizontal and vertical directions of character images. after transforming two-dimensional image into one dimensional signal via projection histogram, then Principal Component Analysis (PCA) is used to distinguish which of the histogram are unique for the particular character and also PCA used to reduce the dimension of features. After feature extracted using projection histogram and features are reduced using PCA. Finally, this reduced feature fed to classification algorithm. In this paper, SVM is used as a classifier. The difference kernels of SVM are used. Hence, RBF kernel achieved good result.

Paper [1] describes Handwritten Gurmukhi character recognition. In this paper there are two features of characters are computed: gradient feature and curvature feature of character image. The two features are fused together to create a single feature vectore.In feature extraction, there are four steps: Computation of gradient, computation of curvature, composite feature vector generation and at the end reducing the dimension of feature using Principal Component Analysis(PCA). There are two feature fusion techniques are used to fuse the two features: forming composite feature by simple concatenation and forming composite feature by cross product. Finally, the combined feature used by the classification algorithm. SVM with RBF kernel used as a classifier. The accuracy of recognition rate is 98.56%.

In paper [7], proposed SVM based offline handwritten digit recognition. In this paper SVM is used as a classifier. The authors performed preprocessing of character image such as binarization, slant correction, smoothing and noise removal and normalization. In feature extraction, the authors used four sets of feature extraction techniques such as using boxing approach, using diagonal distance approach, mean and gradient operations. Enormous features are extracted using those techniques. The extracted features are concatenated to form a single feature vector. After obtaining these features, SVM is applied to the final feature vector. The accuracy of recognition rate is 97.16%.

III. PROPOSED SYSTEM AND METHODOLOGY

The proposed system consists of different methodologies and techniques. Some of the techniques which are discussed subsequently are image acquisition, preprocessing, feature extraction and classification and recognition. Our work mainly focusses on isolated Amharic handwritten character recognition. The proposed system depicted graphically in fig.3 below.

3.1 Image acquisition

Image acquisition is the technique of acquiring images using scanner and digital camera. The collected image format might be JPEG, GIF, PNG and BMP. All the image should be having the same format. After acquiring the image, the image used as input for preprocessing step for further analysis. After collecting the characters written on paper, the paper is scanned using Epson scanner with 300dpi (dot per inch) resolution.

3.2 Preprocessing

Input data should be preprocessed for further analysis. Scanned character image used as input for preprocessing technique. Preprocessing is the basic step in character recognition. The aim of preprocessing is to remove noisy, and to make simple the subsequent task. It includes gray scaling, smoothing and noise removal, normalization, dilation and filling, and the processed image used as input for the segmentation phase [8]. Since the scanned image contain noise and the size of the scanned image is not normalized. In our study, preprocessing is performed on scanned image using image processing techniques and image processing MATLAB toolboxes.

3.3 Feature extraction

Feature extraction is the process of selecting or extracting the relevant features from the image. In handwritten character recognition feature extraction is the most important task and used to achieve high performance of recognition [13]. In character

recognition, the features that uniquely identify the character is described in feature vector. Extracted feature vector used as input to the classification and recognition algorithm. Based on the feature, the classification algorithm recognizes the character. In general, accurate and distinguishable feature plays a significant role to leverage the performance of a classifier [15]. In other words, the character's feature extracted in proper way, the classification algorithm recognizes the character effectively. For our study, 2 D haar discrete wavelet transform followed by histogram of oriented gradients are proposed.

3.3.1 Two Dimensional Haar Wavelet

Wavelet is a function that waves above and below the x-axis with varying frequency, limited duration, zero average value. It is a type of global transformation and series expansion feature extraction technique. There are some basic properties of wavelet transform that is distinguished from Fourier transform: Simultaneous localization in time and scale, sparsity, adaptability and linear time complexity. Haar wavelet introduced in 1910 by Alterd Harr, it is called compact, dyadic and orthogonal wavelet transform. The best features of haar is it provides a local analysis of signals. Haar has different application: image coding, edge extraction and binary logic design [16,22,36]. It is simplest and discrete types of wavelet transform. One dimensional haar wavelet is computed by averaging and subsampling the pixel together(pairwise) to get a new lower resolution image.

Two-dimensional harr wavelet transform is projection of an image onto two-dimensional harr basis functions, obtained by the tensor product of the one-dimensional harr scaling and wavelet functions [16,22,36]. Two-dimensional harr wavelet decomposition can be calculated from one dimensional haar wavelet decompositions. The two different decompositions such as standard decomposition and non-standard decomposition. In standard decomposition, first compute one dimensional haar wavelet decomposition of each row of the original pixel values. Next compute one dimensional decomposition of each column of the row-transformed pixels. Whereas non-standard decomposition, alternates between operation on rows and columns. First perform one level decomposition in each row (i.e., one step of horizontal pairwise averaging and differencing). Second perform one level decomposition in each column from step 1 (i.e., one step of vertical pairwise averaging and differencing). Rearrange terms and repeat the process on the quadrant containing the averages only.

Haar wavelet mother function, $\Psi(t)$ can be described as [16,22,36],

$$\psi(t) = \begin{cases} 1 & 0 \le t < \frac{1}{2}, \\ -1 & \frac{1}{2} \le t < 1, \\ 0 & otherwise \end{cases} \tag{3.1}$$

Its scale function $\varphi(t)$ can be defined as [16,22,36],

$$\varphi(t) = \begin{cases} 1 & 0 \le t < 1, \\ 0 & otherwise \end{cases} \tag{3.2}$$

A family of functions can be obtained from the basic by scaling and translation (compress and stretch, respectively),

$$\varphi_{i,j}(t) = 2^{i/2}\varphi(2^i - j) \tag{3.3}$$

Scaling function, $\varphi_{i,j}(t)$ span in the vector space V^i, nested as $V^0 \subset V^1 \subset V^2 \subset \ldots$

3.3.2 Histogram of Oriented Gradients (HOG)

Histogram of oriented gradients are feature descriptor used to extract the features of an image by counting the occurrence of gradient orientation in localized parts of an image [32]. Preforming feature extraction using HOG, image is partition into cells and histogram of gradient are formed from each cell. In HOG feature extraction, it easy to express the rough structure of the object and is robust to difference in geometry and illumination changes. HOG feature extraction method used in different application, such as human detection [32], pedestrian detection [33], character recognition [34], baggage detection [31] and so on. HOG feature has several advantages: It capture edge or gradient structure which is very characteristics of local shape and does local representation with an easy controllable degree of invariance to local geometric and photometric transformations, which means translations and rotations make little variance if it is much smaller that the local spatial or orientation bin size [32]. HOG feature extractor used with specified matrix of pixels, which slides over the entire image. In each position of the image, HOG descriptor is computed.

Generally, HOG algorithm works as follow:

(i) Divide the character image into connected region, known as cells, and compute gradient orientation and gradient magnitude of each pixel in each cell.
(ii) Discretized each cell into histogram bins based on the gradient orientation.
(iii) Each cell's pixel produces weighted gradient to the equivalent bin.
(iv) Group adjacent cells into blocks of size 16x16 pixels
(v) Normalized group of blocks using L2-Norm, finally, these blocks are used for feature descriptor.

Four coefficients are obtained from original character image using 2-dimensional harr wavelet. After obtained these harr coefficients, histogram of oriented gradients in the four coefficients is extracted. Finally, concatenating all features together in order to create one feature vector. The dimension of the feature is large. In order to reduce the dimension of feature, Linear Discriminant Analysis (LDA) is applied. LDA is used to reduce the dimension. LDA project the feature in smaller dimension. It has discriminative power of differentiating classes by keeping the classes structure. It is used in supervised machine learning algorithm. The reduced feature vector used as input for classification algorithm.

3.4 Classification and Recognition

Classification takes place after feature extraction. Classification is the process of discriminating one class from the rest of classes based on different criteria. In other words, classification methods are used to apply in pattern recognition to separate one class from the rest of classes. At this stage, characters can be recognized based on their features. The decision has been made in this step. Different classification algorithms used depend on their performance and accuracy: Artificial neural network (ANN), Support vector machine (SVM) [14], soft max layer in case of CNN [5,6], deep belief network [3], adaboost, decision tree, genetic algorithm [4], KNN and so on.

3.4.1 Support Vector Machine (SVM)

Support vector machine is a supervised machine learning algorithm used to solve linear and nonlinear classification problem [30]. SVM is based on the structural risk minimization, it is better than empirical risk minimization used in neural network [30]. The aim of SVM is to find a parameter setting that decrease the risk famulated by,

$$R_{emp}(\alpha) = \frac{1}{2l}\sum_{i=1}^{l} | y_i - f(x_i, \alpha) | \tag{3.10}$$

Where y_i represent the output, x_i 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 distance, margin. The best characteristics of support vector machine is robust to outlier. SVM find a global minimum [9]. There are two cases in which SVM is used. The first case is separable [9]. Label data such as $\{xi, yi\}, i = 1,2,...l, yi \in \{-1,1\}, xi \in R^d$ [26]. 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 [9]:

$$x_i.w + b \geq +1 \quad for \quad y_i = +1 \tag{3.11}$$

$$x_i.w + b \leq -1 \quad for \quad y_i = -1 \tag{3.12}$$

The combination of the two equation is

$$x_i.w + b - 1 \geq 0 \quad \forall_i \tag{3.13}$$

Consider equation (3.13), the points lie on hyperplane h1: x_i.w+b=1 with normal w, perpendicular distance is |1-b|/||w|| and in equation (3.14), the points lie on hyperplane h2:x_i.w+b=-1 with norm w and the distance from the origin |-1-b|/|b||. In general, SVM find hyperplanes that gives largest margin by maximizing $\| w \|^2$, based on the equation (3.14) [9].

$$w = \sum_i \alpha_i y_i x_i \tag{3.14}$$

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

The Lagrangian formulation the optimization problem becomes,

$$L = \frac{1}{2} || w ||^2 - \sum \alpha_i y_i (w^T x_i + b - 1) \quad a_i \geq 0 \tag{3.15}$$

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, …, l, so the linear separable equation changed to the following equation:

$$y_i(x_i.w + b) \geq 1 - \xi_i \ for \ i = 1,2,..l \tag{3.16}$$

The non-separable problem becomes,

$$\text{Min } \frac{1}{2} || w ||^2 + C\sum_i \xi i \tag{3.17}$$

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 equation (3.18) [9].

$$w = \sum_i \alpha_i y_i x_i \tag{3.18}$$

In this case, the difference between equation (3.14) and equation (3.18) is that α_i have an upper bound of C.

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

$$\Phi : R^d \rightarrow H \tag{3.19}$$

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 space, $k(x_i, x_j) = \Phi(x_i).\Phi(x_j)$ .where k is representing kernel function. There are tuning parameters in SVM such as kernel, regularization, gamma and margin. Linear kernel is used in this work. The most common Kernel functions are:

Linear kernel: $k(x_i, x_j) = x_i \times x_j$

Polynomial kernel: $k(x_i, x_j) = [(x_i \times x_j) + 1]^d$

Sigmoid kernel: $k(x_i, x_j) = \tanh(\beta_0 x_i x_j + \beta_1)$

Extensional Radial Basis Function Kernel (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 equation (3.20),

$$f(x) = \sum_{i=1}^{Ns} \alpha_i y_i \Phi(s_i).\Phi(x) + b = \sum_{i=1}^{Ns} \alpha_i y_i K(s_i, x) + b \tag{3.20}$$

Where Si are support vectors.

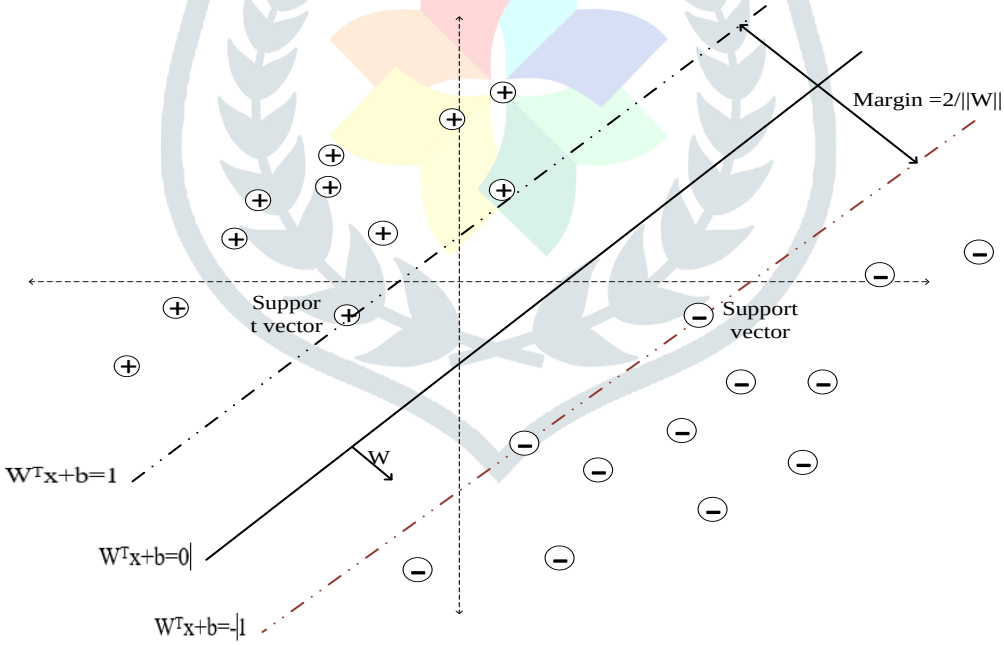

Fig. 2. Binary class SVM

Binary classification problems have been studied in different literatures. And also, multiclass classification has been introduced to solve multiclass 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 support vector machine is an 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) [9], one versus one(OVO) [10] and Error Correcting Output Code(ECOC) [10] [25]. 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 [10]. 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 [27]. 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 all has N bit codewords [27]. One versus one strategy considers all possible pairs of classes [27]. Its codeword length could be N(N-1)/2. Binary classifier consists of group of independent binary classifiers are trained based on different categories of the original data, based on each column of the coding matrix. Whereas decoding is the final classification is 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.

3.5 Feature Reduction using Linear Discriminant Analysis(LDA)

Linear discriminate analysis is a technique used in pattern recognition and machine learning to find a linear combination of features that characterize or separate two or more classes of objects or events. LDA is used to express one dependent variable as a linear combination of other features or measurements. Unlike principal component analysis, LDA is used to attempts the model the difference between the classes of data. It is used when groups are a prior unlike cluster analysis. Each case must have one or more quantitative prediction measure, and a score on a group measure. LDA is used commonly as a dimensional reduction method in the preprocessing steps for pattern-classification and machine learning. It projects the data set onto a lower dimensional space with good class separability in order to avoid overfitting. And also, computational cost. In contrast to principal component analysis, LDA is supervised and compute the direction (linear discriminants) that will represent the axes that maximize the separation between multiple classes [11,35].

For c classes, the label class i as x_i and denote by L_i the number of training character image in x_i. The mean vector for class x_i is m_i and over all mean vector for character image is m, where $\mathbf{L} = \sum_{i=1}^{c} \boldsymbol{L}_i$.

Between class scatter matrix is given as,

$$S_{\mathrm{B}} = \sum_{i=1}^{c} L_i (m_i - m)(m_i - m)^T \tag{3.24}$$

Within-class scatter matrix also given as,

$$S_{\mathrm{w}} = \sum_{i=1}^{c} \sum_{x_j \in x_i} L_i (x_j - m_i)(x_i - m_i)^T \tag{3.25}$$

If S_w is nonsingular, then the optimal projection W_{opt}[11,35] is computed as,

$$W_{\mathrm{opt}} = \arg\max_{w} \left| \frac{W^T S_{\mathrm{B}} W}{W^T S_{\mathrm{w}} W} \right| = [W_1, W_2, \ldots, W_m] \tag{3.26}$$

Where $\{W_i | i\text{-}1,2,\ldots,m\}$ is a set of generalized eigenvectors of two scatter matrices. The upper bound of mean m is c-1, the maximum rank of S_B. And detail explanation of W_{opt} is given in [11,35].

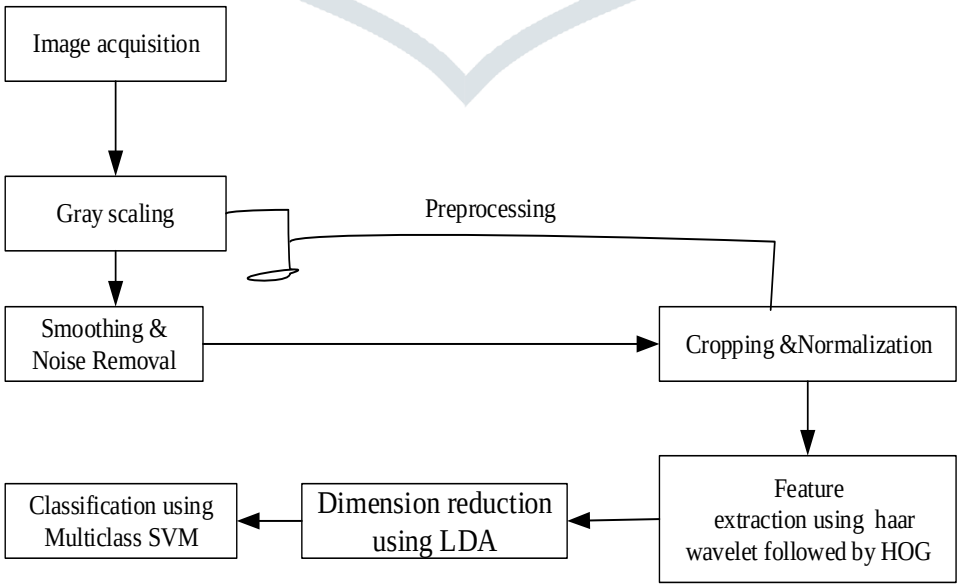

Fig. 3. Amharic handwritten character recognition flow chart diagram

3.6 Data Set Preparation

Data set is necessary for developing optical character recognition system. Data set is used to train and testing algorithm. Particularly, in handwritten character recognition, large amount of data set is needed train the algorithm. Since different persons have different writing style. Large amount of data from different persons based on distinct group of people are needed. Some of non-Latin characters such as Amharic characters have no data set available online. Collecting from scratch and building handwritten character corpus is tedious and boring. Amharic handwritten character data set (AHCD) collected from scratch based on various groups of people. We prepared grid like A4 papers for each character. Different writing styles from different persons are collected. The data set was collected from three categories: In first category, AHCD collected from post graduate students. In the second category, AHCD collected from undergraduate students. In the third category, AHCD collected from non-native speakers and native speakers of Amharic language. Sample data set for the first fourteen Amharic handwritten characters written by five different persons are shown below in fig 4. After collecting the dataset, it cropped out the individual character. In the cropped character data set, there is white spaces at the top, bottom, left and right of the characters. The white spaces from four directions are removed, such as top, bottom, left and right. The character images resized to 64 wide with 128 tall by keeping their aspect ratio. Our dataset is important to perform further research in Amharic handwritten characters. and also, Chars74K benchmark numeric dataset [29] is used to validate the proposed approach. It consists of Latin script and Hindi-Arabic numerals.

Fig. 4. Sample data set for the first fourteen Amharic handwritten characters

IV. Experiment and Result

To implement the algorithm, 8GB RAM, intel core i7 processor and 2TB hard disk dell laptop computer are used. And MATLAB programming language used to implement and simulate algorithm. For this work MATLAB 2017a version is used. The proposed model trained and tested based on different parameters of HOG and support vector machine. For HOG, 4x4 pixels of cell size, 16x16 pixels of block size, and L2-norm to normalize histogram in overlapping and nonoverlapping blocks are used. In each block the histogram is normalized to eliminate shadow, illumination and contrast. For calculating gradient magnitude and orientation, sobel operator is used, and unsigned degree, 0^0-180^0 for orientation of 9 bins. Different kernel functions of multiclass SVM are implemented. Both gaussian kernel and polynomial kernel recorded low accuracy compared to linear kernel by changing gamma and order, respectively. Linear kernel produced good result for Amharic handwritten character recognition. In addition to using different kernel function, Multiclass SVM algorithm using one versus one and one versus all design coding matrix is implmented. As a result, one versus all better than one versus one design coding matrix for Amharic handwritten character recognition and Chars74K benchmark numeric data set. One versus one is computationally expensive. The model is validated using 10-fold cross-valuation method. 10-fold cross-validation is one of good model validation techniques by dividing the data set into 10 folds. From the result obtained, some of the character gave good result. However, characters which have visual similarity of shape have given low result. Characters, such as ፈ, ፉ, ፊ, ፋ, ፌ, ፍ, ፎ, ጪ, ጭ, ፐ, ፕ, ፖ, ሸ, ሽ, ም, ሞ, ረ, ሩ, ሪ, ራ, ሬ, ር, ሮ, ፀ, ጸ, ቈ, ዮ and ዯ are produce low accuracy than other characters. These characters are difficult to recognize even for human being. Amharic characters are large in number to represent using graphically and using confusion matrix. Confusion matrix is obtained using MATLAB and the accuracy is summarized using table 1 below implemented on AHCD data set. And also, proposed approach applied on Chars74K benchmark numeric data set and results are shown in table 2 below. In general, the results obtained by applying the proposed approach on AHCD data set and Chars74K benchmark numeric data set are shown table 1 and table 2 below, respectively, obtained from the confusion matrix. As we observed from the result obtained, the accuracy of recognition using haar wavelet followed by HOG better than HOG alone. LDA also helps to reduce the dimension of feature vector into lower dimensional feature. LDA has good features in

discriminating classes. When the proposed approach applied on digit handwritten characters, it gave us a good result. Digit has ten classes, so it is simple to show using confusion matrix. confusion matrix of digit recognition using the proposed approach is described in fig. 5 below. The diagonal elements of the confusion matrix show the correct classification of each character. The last diagonal element describes the overall accuracy of recognition. As we stated visual similarity of Amharic characters still the challenging. It reduces the overall accuracy of recognition.

4.1 Result of Proposed and Existing System

Both existing and the proposed system are implemented. In existing system, the basic OCR techniques like preprocessing, feature extraction and classification are used. Specially in feature extraction, Histogram oriented gradients feature descriptor is used. HOG feature computed in 128x64 gray scale image size. In this technique, sobel operator is used by convolving the whole character image to compute gradient magnitude and gradient orientation. In HOG feature extraction, 4x4 pixel cell size , 16x16 pixel block size, 0^0-180^0 bin orientations of nine bins (20^0 orientation of each bin) are used. The features in histogram are normalized using L2-norm in order to remove shadow, illumination and contrast. After extracting HOG feature on gray scale image, SVN is used as a classifier. Existing approach have been trained and tested on AHCD Amharic data set and Chars74K benchmark numeric data set. The existing system achieved accuracy of 82.90% on AHCD data set and 91.25% on Chars74K benchmark numeric data set.

Based on the existing approach, Harr wavelet transform and HOG are proposed on Amharic handwritten characters and Chars74K benchmark numeric data set. In this approach, first harr wavelet transforms applied on gray scale 128x64 image size. Harr wavelet transform decompose the given grayscale image into four coefficients, such as average, horizontal, diagonal and vertical coefficient. In each coefficient, HOG features are extracted. The extracted HOG feature from each coefficient are concatenated to form single feature vector. In this approach, using harr wavelet edge of the character image is detected like horizontal and vertical. This approach helps to reduce the size of the feature and extract the important features from four coefficients rather than extracting from single image. Similarly, used 4x4 pixel cell size, 16x16 pixel size, 0^0-180^0 unsigned degree and 9 bines are used. Finally, LDA is applied to the feature vector to reduce the dimension of the feature. The final feature vector fed to the classification algorithm called multiclass SVM. This approach applied to AHCD data set and Chars74K benchmark numeric data set. It achieved an accuracy of recognition 86.30% on AHCD data seta and 93.10% on Chars74K benchmark numeric data set.

TABLE 1 ACCURACY OF RECOGNITION USING PROPOSED APPROACH ON AMHARIC HANDWRITTEN CHARACTER DATA SET

Sr.no	Feature Used	Classifier	Accuracy
1	HOG	linear SVM	82.90%
2	**Haar Wavelet and HOG**	**linear SVM**	**86.30%**

TABLE 2 ACCURACY OF RECOGNITION USING PROPOSED APPROACH ON CHARS74K BENCHMARK NUMERIC DATA SET

Sr.no	Feature Used	Classifier	Accuracy
1	HOG	linear SVM	91.25%
2	**haar Wavelet and HOG**	**linear SVM**	**93.10%**

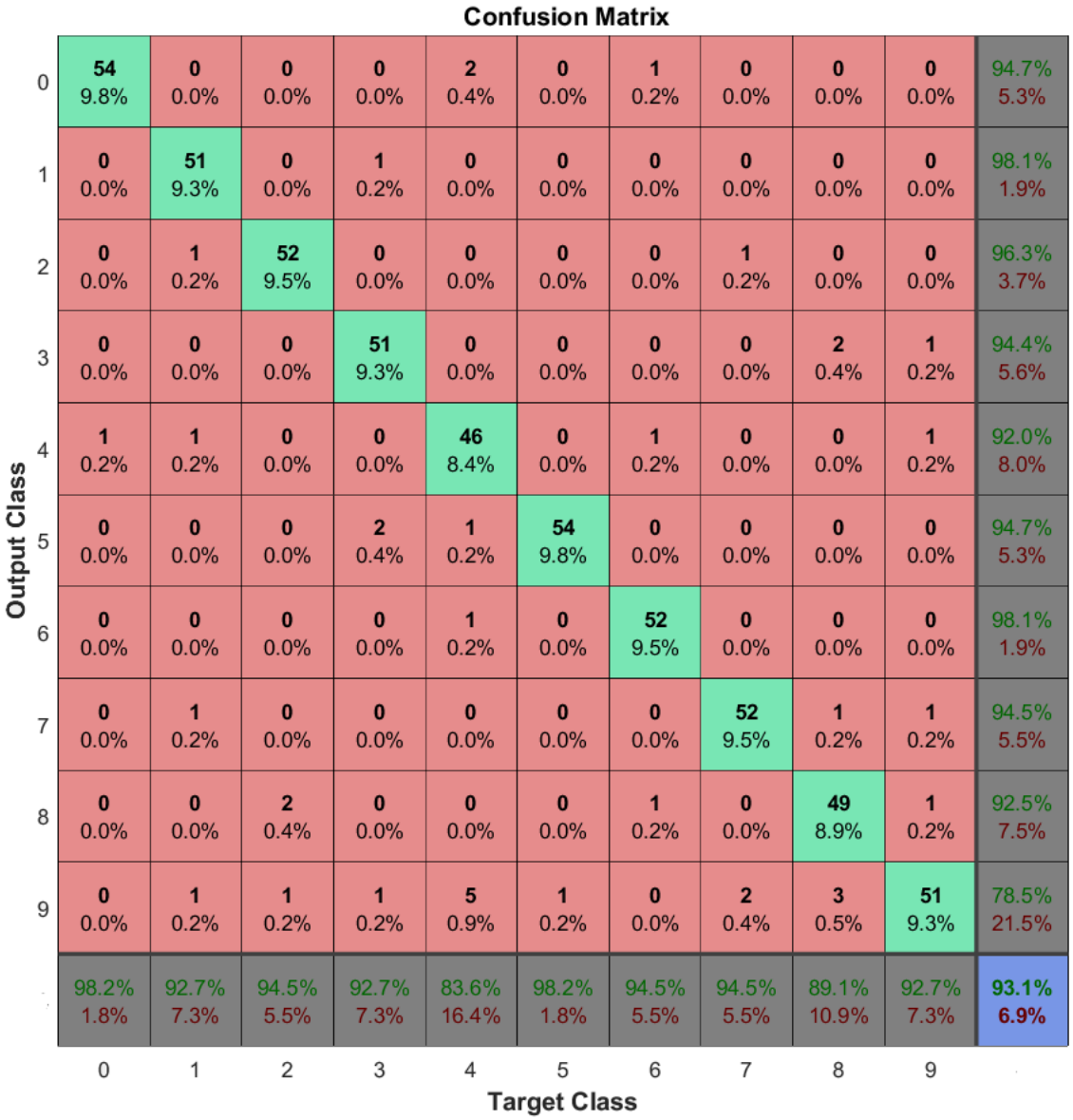

Fig. 5. Confusion matrix for the accuracy of digit recognition using proposed approach

V. CONCLUSION

Handwritten character recognition is the most challenging task in pattern recognition. Since handwriting of one person is differ from another person. In addition to variability of different writing styles, in Amharic characters, there is visual similarity of 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), which makes the recognition task difficult. Different papers have been reviewed related to our study, and problem of the system and its challenge are identified. The different techniques like image preprocessing, feature extraction and classification discussed. Feature extraction is the most important method which determine the accuracy of the recognition algorithm. Character features are extracted using Haar wavelet transform followed by HOG. The dimension the feature is reduced of using Linear Discriminant Analysis(LDA). Experiment conducted on Amharic characters and digit data set using HOG and haar wavelet followed by multiclass SVM classifier as classification algorithm. And the model evaluated using 10-told cross-validation technique. From the result shown most of the characters classified correctly and some of the characters are misclassified. In future work, we suggest that using deep learning the accuracy will be improved by increasing the data set. And also, using another feature extraction techniques the accuracy will be improved. Working on large data set, cursive writing, special symbol, Amharic digits are future work. Visual similarity of character's shape is still challenging. It is also future work.

REFERENCES

[1] A. Aggarwal and K. Singh, "Handwritten Gurmukhi character recognition," 2015 International Conference on Computer, Communication and Control (IC4), Indore, 2015, 1-5

[2] El Moubtahij, Halli and Satori, "Recognition of Off Line Arabic Handwriting Words Using Hidden Markova Model(HMM) Toolki," 13th International Conference on Computer Graphics, Imaging and Visualization (CGiV), Beni Mellal, 2016, 167-171.

[3] M. R. Sazale, S. K. Biswase, M. F. Amein and K. Muraese "Bangla handwritten character recognition using deep belief network,"International Conference on Electrical Information and Communication Technology (EICT), Khulna, 2014, 1-5.

[4] T. Sahlole, Y. Suen, M. Zawba, A. E. Hassanien and A. Elfattah, Bio inspired BAT optimization algorithm for handwritten Arabic characters recognition," IEEE Conger. On Evolutionary Computation (CEC), Vancouver, BC, 2016, 1749-1756.

[5] S. Yang, F. Nian and T. Li, "A light and discriminative deep networks for off line handwritten Chinese character recognition," 32nd Youth Academic Annu. Conference of Chinese Association of Automation (YAC), Hefei, 2017, 785-790.

[6] P. Naire, A. Jamese and C. Saravanane, "Malayalam handwritten character recognition using CNN," International Conference on Inventive Communication and Computational Technologies (ICICCT), Coimbatore, 2017, 278-281.

[7] G. Katiyar and S. Mehfuz, "SVM based off-line handwritten digit recognition," 2015 Annual IEEE India Conference (INDICON), New Delhi, 2015, 1-5

[8] A. Prieya, S. Mishrae, S. Raj, S. Mandale and S. Datta, "Online and off-line character recognition: A survey," 2016 Internationall Conference on Communication and Signal Processing (ICCSP), Melmaruvathur, 2016, 0967-0970.

[9] A. Mowlaei and K. Faez, "Recognition of isolated handwritten Persian/Arabic characters and numerals using support vector machines," 2003, 547-554.

[10] Mohammad ali Bagheri," Error Correcting Output Codes for multiclass classification: Application to two image vision problems" CSI International Symposium on Artificial Intelligence and Signal Processing, Barcelona, Spain ,2012.

[11] Kale, Amit C., and R. Aravind. "Face recognition using canonical correlation analysis." In the National Conference on Communications, 2007, 48-52.

[12] A. Rosenfeld and A. Kak In Digital Picture Processing; First edition; Academic Press, New York, 1976.

[13] 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, 2014, Vol. 3, No.1 , 481–485.

[14] R. Li et al., "Feature Extraction and Identification of Handwritten Characters," 8th International Conference on Intelligent Networks and Intelligent Systems (ICINIS), Tianjin, 2015, 193-196.

[15] K. L. Kabir "A superior domain for handwritten Bangla basic characters recognition," IEEE 9th International Conference on Intelligent Systems and Control (ISCO), Coimbatore, 2015, 1-7.

[16] P. Porwik and A. Lisowska, "The haar wavelet transform in digital image processing: Its status and achievements," Machine graphics & vision,2004, vol. 13, no. 1, 79–98.

[17] J. Cowell and F. Hussain, "Amharic character recognition using a fast signature based algorithm," Procceding on Seventh International Conference on Information Visualization, 2003, 384-389.

[18] W. Alemu and S. Fuchs, "Handwritten Amharic Bank Check Recognition Using HMM Random Field," Conference on Computer Vision and Pattern Recognition Workshop, Madison, Wisconsin, USA, 2003, 28-28.

[19] Y. Assabie and J. Bigun, "Lexicon based offline recognition of Amharic words in unconstrained handwritten text," 19th International Conference on Pattern Recognition, Tampa, FL, 2008.

[20] Y. Assabie and J. Bigun, "HMM Based Handwritten Amharic Word Recognition with Feature Concatenation," 10th International Conference on Document Analysis and Recognition, Barcelona, 2009, 961-965.

[21] C. Cheng, Y. Zhange, X. H. Shaou and D. Zhoue, "Handwritten Chinese Character Recognition by Joint Classification and Similarity Ranking," 15th International Conference on Frontiers in Handwriting Recognition (ICFHR), Shenzhen, 2016, 507-511.

[22] C. Burrus, R. Gopinath, and H. Guo, Introduction to wavelets and wavelet transforms: A Primer. Prentice-Hall, 1998.

[23] H. Lee et al., "Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations," In Procceding of the 26th Annu. International Conference on Machine Learning, ACM, 2009, 609–616.

[24] A. Youan, G. Baie, L. Jieao and Y. Liue, "Off-line handwritten English character recognition based on CNN," 10th IAPR International Workshop on Document Analysis Systems, Gold Cost, QLD, 2012, 125-129.

[25] Thomas G.Dietterich and Ghulum Bakiri" Solving Multiclass Learning Problems via Error-Correcting Output Codes," journal of Artificial Intelligence Research 2, USA ,1995,263-286.

[26] M.W. Gardner, S.R. Dorling," Artificial Neural Networks-MLP", Elsevier Science in Atmospheric Environment, Greater Britain, 1998, 2627-2636.

[27] S. Escalera, O. Pujol and P. Radeva, "On the Decoding Process in Ternary Error-Correcting Output Codes," in IEEE Transactions on Pattern Analysis and Machine Intelligence, Jan 2010, vol. 32, no. 1, 120-134.

[28] L. Pauly, "Hand written digit recognition system for South Indian languages using ANN," Eighth International Conference on Contemporary Computing (IC3), Noida, 2015, 122-126.

[29] 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.

[30] 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, Cebu City, 2015, 1-6.

[31] T. Khanam,"Baggage detection and classification using human body parameter & boosting technique," 10th International Conference on Human System Interactions (HSI), Ulsan, 2017, 54-59.
[32] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), San Diego, CA, USA, 2005, 886-893.
[33] D. Tasson, "FPGA based pedestrian detection under strong distortions", Computer Vision and Pattern Recognition Workshops (CVPRW) IEEE Conference on, 2015, 65-70.
[34] O. L. Junior, "An application to pedestrian detection," 12th International IEEE Conference on Intelligent Transportation Systems, St. Louis, MO, 2009, 1-6.
[35] Sebastian Raschka," Linear Discriminant Analysis," April, 2018 http://sebastianraschka.com/Articles/2014_python_lda.html.
[36] S. Banerji, A. Sinha and C. Liu, "HaarHOG: Improving the HOG Descriptor for Image Classification," 2013 IEEE International Conference on Systems, Man, and Cybernetics, Manchester, 2013, pp. 4276-4281.