ROBUST FEATURE-BASED RECOGNITION OF PRINTED COMPREHENSIVE HAUSA CHARACTERS

Tunji S. Ibiyemi [1], Yakubu A. Ibrahim [2]

[1]Department of Electrical Engineering, University of Ilorin, Ilorin, Nigeria
[2]Department of Computer Science, Bingham University, Karu, Nasarawa State, Nigeria

Corresponding author: Tunji S. Ibiyemi, ibiyemits@yahoo.com,
Yakubu A. Ibrahim, talktoibro80@gmail.com

ABSTRACT: *Rapid growth of technology and prevalent use of computer in the business and other areas appeals for technology advancement, most organizations are converting their paper documents into electronic documents that can be processed by computer. Much research has been done on the identification of English, Arabic, Japanese and Chinese. It is observed that the research on recognition of printed characters for most of the African languages including Hausa is still an open research problem. However, Hausa language is one of the dominant languages in sub-Sahara Africa hence the study. Here, the study considers the processing of gray level images only, since they contain enough information to perform feature extraction and image analysis. Therefore, the paper develops a simple and efficient method for the recognition of isolated printed Hausa characters. Images of printed Hausa characters were captured using a scanner and images were pre-processed to remove noise. The methodology for isolated printed Hausa character recognition was based on efficient feature extraction followed by a suitable feature vector dimensionality reduction scheme. HMM algorithm was adopted to develop a system for recognition of printed Hausa characters.*

KEYWORDS: *DTW, Hausa Language, HMM, Speech Recognition, MFCC.*

1. INTRODUCTION

Pattern recognition is fast moving research area. The advent of powerful OCR computing devices triggered statistical methods such as HMM to improve robustness of the pattern classifier across various printed characters. Printed characters can be written in a various number of ways using different shapes and properties. As a matter of fact, printed character recognition in recent years has been a challenging and fascinating research area in field of image processing [3]. Advancement in pattern recognition and automation process can improve human and computer interaction in many areas of life. Many research works in pattern recognition nowadays focuses on new techniques that would minimize the processing time while giving higher system recognition accuracy [5]. In general, a widely used approach in isolated character recognition follows a two step procedures which represent the character as a vector of features and group the feature vectors into different classes of objects. Choosing an appropriate feature extraction method is very vital in developing a high pattern recognition system. Feature extraction technique should be robust to an extent that for a different of instances of the same symbol, similar feature vectors are produced, hence, making the subsequent grouping or classification task not only less difficult but much easier [4].

Many computer applications including document reading, bank processing, mail sorting, and postal address recognition require a printed character recognition systems. As a result, the printed characters recognition system is always an active area for research towards discovering the efficient methods that would improve pattern recognition accuracy [6].

2. OPTICAL CHARACTER RECOGNTION

The Two-Dimension digital image a[m,n] has two basic components which are N rows and M columns. The intersection point of row and column is called picture element or a pixel. The value allocated to the integer coordinates [m,n] with {m=0,1,2,…,M–1} and {n=0,1,2,…,N–1} is a[m,n]. A computer image is a matrix (a two-dimensional array) of pixels. The value of each pixel is proportional to the brightness of the corresponding point in the scene; its value is usually derived from the output of an analogue-to-digital (A/D) converter [7].

3. TYPES OF DIGITAL IMAGES

a. Binary Image: In this type of image each pixel of the image contains black or white. The two possible colours generate the two possible values for each pixel in a given image. Binary images in other words can be very efficient in terms of storage.

b. Grayscale Image: Each picture element of a grayscale image is a shade of gray, usually from Zero (black) to Two Hundred and Fifty Five (white). This range of the above type of image means that each picture element can be represented by eight bits or one byte.
c. True Colour or RGB Image: In this type of image each picture element has a particular colour; that colour being explained by the amount of red, green and blue in it. If each of the three colours has a range from Zero to Two Hundred and Fifty Five (0-255), then, it gives a total of 2553=16, 777, 216 different possible colours in the digital image. This variety of different possible colours is enough for any digital colour image.
d. Indexed Image: Almost all the colour digital images only have a small part of the more than sixteen million possible colours. For proper storage and effective file handling, the image has an associated colour map which is simply a list of all the colours used in that image. Each picture element has a value which does not give its colour (as for an RGB image), but an index to the colour in the colour map.

4. THE PROPOSED PRINTED HAUSA CHARACTER RECOGNITION SYSTEM

In this section, the proposed recognition system is described. A typical printed character recognition system consists of pre-processing, segmentation, feature extraction, matching/classification and recognition stages. The schematic diagram of the proposed recognition system is shown in Figure 1.

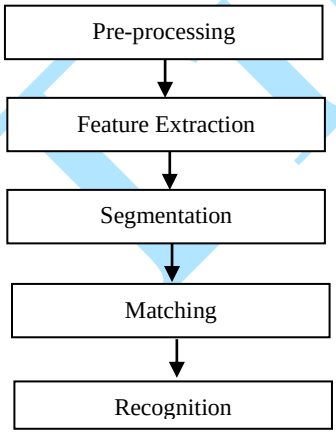

Figure 1: Block diagram of the proposed printed Hausa recognition system

5. DATA SET DESCRIPTION

The availability of data set that captures variations encountered in real world is a critical part of any experimental research. Due to significant advances in the OCR research in recent years, several data sets are available for English. However, no printed Hausa character data set exists. Therefore, sample of each printed Hausa characters was developed and scanned to form a data set. A total of 96 Hausa character graphical images were designed in eight-by-fifteen grids and their bitmap created on computer to form Hausa character data-set. These characters are: 60 consonants, 10 vowels, 10 digits, and 16 special characters. Thus this data set consists of 125 samples of Hausa printed characters altogether [1].

A. Data/Image Acquisition

Optical scanners are often employed in OCR systems to obtain digital documents. The input image is threshold to generate a binary image. In a binary image, gray levels below a selected threshold value will be treated as black, and the values above threshold are treated as white. Thresholding process saves memory as well as computational efforts [10].

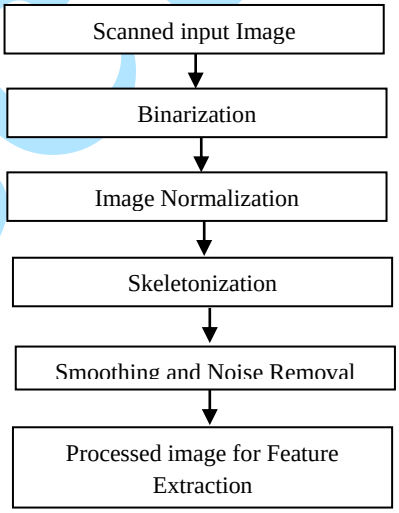

Figure 2: Shematic diagram of Pre-processing

In the study, the pattern recognition system uses a scanned digital image as an input image. The digital image could be in a specific format of JPEG, PNG, BMP, BMT etc. The images were acquired in PNG file, but can as well be obtained via a camera or any other input device.

B. Pre-processing of Image

The pre-processing of an image is a sequence of operations which could be performed on the scanned input digital image. It basically enhances the digital image rendering which is suitable for image segmentation. The different work done on the any digital image at pre-processing stage is shown in Figure 2.

1. Binarization: is to convert the scanned image into a binary image. However, binarization is an

operation that produces two classes of pixels, in general, they are represented by black pixels and white pixels. The process changes a gray scale digital image into a binary digital image adopting the global thresholding technique.

2. Normalization of the position

The goal was normalization is to improve the processing time by removing unwanted areas. In order to evaluate the shape of a signal, regardless of its overall amplitude, the image has to be normalized. Hence, normalization has to take three aspects of the signal into consideration:

a. Offset: to normalize the offset, one can subtract the mean of the signal, or the smallest value of the signal. (The latter option ensures that the resulting signal is always positive).
b. Duration: to ensure that two images have the same length, one can interpolate such images to a fixed number of data points.
c. Amplitude: one way to normalize the amplitude is to scale the signal such that the smallest value is 0, and the largest value is 1. However, a better way is to normalize it such that the maximum value of the auto-correlation function is equals to one [4]. In this regard, one calculate the horizontal and vertical histograms to detect the first white pixel at top, bottom, left, right.

3. Skeletonization:

The basic idea of skeletonization is to reduce the thickness character image to one-pixel while preserving its connectivity and its topological properties. The skeleton must preserve the shape, connectivity, topology and end of the route, and should not introduce parasitic elements. Thinning which is the morphological operation removes the highlighted foreground pixels from binary images, somewhat like erosion or opening. It is particularly used along with skeletonization.

These two operations are performed for the purpose of extraction of pattern descriptor feature [9]. The major skeletonization steps are:

a. Detecting ridges in distance map of the boundary points,
b. Calculating the Voronoi diagram generated by the boundary points, and
c. The layer by layer erosion called thinning.

In digital spaces, only an approximation to the true skeleton can be extracted. There are two requirements to be complied with:

a. Topological requirement: to retain the topology of the original object,
b. Geometrical requirement: forcing the skeleton being in the middle of the object and invariance under the most important geometrical transformation including translation, rotation, and scaling [9].

4. Smoothing

The erosion and dilation smooth the boundaries of an Image. Sobel method was used to detect edges in the binarized image and holes present in the digital image are filled [2].

5. Noise Removing

Noise which is in the images is one of the big difficulties in optical character recognition process. The aim of this part is to remove and eliminate this difficulties; there are several methods that allow us to overcome this problem. In this work, morphology operations were adopted to detect and delete small areas of less than 40 pixels.

6. Salt and Pepper Noise

This degradation can be introduced by sharp and sudden disturbances in the digital image signal; the presence of this type of noise is known through randomly scattered white or black (or both) pixels over the digital image.

The two methods for cleaning salt and pepper noise are low pass filtering and median filtering:

a. Low pass filtering

When there is presence of salt and pepper noise in pixels such pixels usually have high frequency components of an image, in this case a low-pass filter should reduce them when used.

b. Median filtering

Median filter is generally used to remove salt and pepper noise from any digital image signal. Recall that the median of a set of values is the middle value of those values when they are sorted in ascending or descending order. If there are even numbers of middle values, the median is the mean of the middle two. A median filter is a typical example of a non-linear spatial filter.

7. Image Restoration

Image restoration targets at reduction and removal degradations that may have occurred in process of digital image acquisition. Those degradations that occurred in any acquired digital image may include noise. Image Noise includes errors in the picture elements values, or image optical effects due to out of focus blurring, or blurring as a result of camera movement [11].

C. Feature extraction

In this section necessary features which can be used in identification process is to be extracted from each character. The feature set for each characters should be unique. There are mainly three types of feature extraction techniques in OCR systems: the distribution of points, transformations and structural analysis.

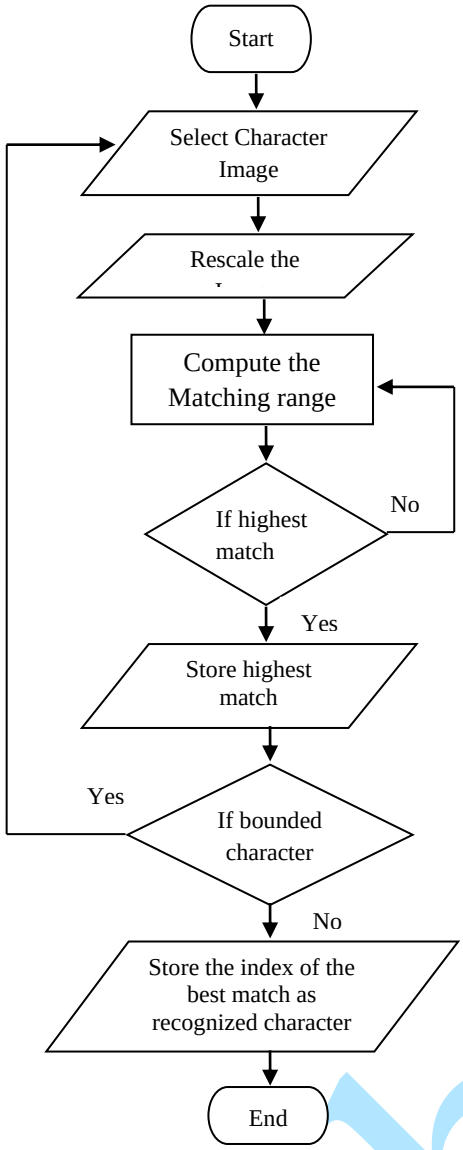

Figure 3: Work flow of the Template matching Algorithm

Distribution of Points Technique:
It extracts features which are based on the statistical distribution of points such as relative joint occurrence of black and white points, crossings and distances in the character shape, moments of black points about a chosen center.

Transformations Technique:
It reduces the dimensionality of the extracted feature set to a great extent. The transformations like Fourier and Hough are usually employed in OCR systems.

Structural analysis Technique:
It extracts features which describes the geometry and structure of a character. This approach is comparatively simple, easy to implement and provides high tolerance against noise and style variations. To some extent, this is still an area of research.

D. Segmentation
This is the process of distinguishing the area in an image where the data is printed. In image processing it is very essential to differentiate the objects of target and the rest objects. The commonly used method to find the objects of target from a given object is called segmentation. It is a technique used for segmenting the foreground of an object from background of an object. The two mostly used techniques are: thresholding and edge detection.
It is vital to know that there is no specific universally applicable segmentation technique that would work effectively for all images in image processing, and, no segmentation method is perfect as well [4]. In OCR systems, usually each character is to be isolated in the segmentation stage. Isolation of characters can be done by extracting each connected black area from a binary image. It is easy to implement but the problem exists when the characters are fragmented.

Thresholding
A grayscale image is turned into a binary (black and white) image by first choosing a gray level threshold (T) in the original image, and then turning every pixel black or white according to whether its grey value is greater than or less than the threshold:
A pixel becomes white if its grey level is greater than T and black if its grey level is less than T. Thresholding is a vital part of image segmentation, where we wish to isolate objects from the background.

Edge detection
Edges contain some of the most useful information in an image. We may use edges to measure the size of objects in an image; to isolate particular objects from their background; to recognize or classify objects.

E. Classification
It is the process of identifying each character from its feature set. Each character should be assigned to its correct class. The commonly used techniques for classification in OCR systems includes: template matching, statistical techniques, structural techniques, and neural networks. In Template matching, the character to be recognized is matched against a set of stored prototypes and according to the degree of similarity, the character will be assigned to a particular class. It is the simplest method of classification. Statistical classifiers are concerned with statistical decision functions and, for using this classifier, sufficient statistics of each class should be available.
Structural approach uses syntactic methods for classification. In syntactic methods, each class will

have its own grammar explaining the composition of a character. Then a class will be assigned to a character, if it can be generated using the grammar of that particular class.

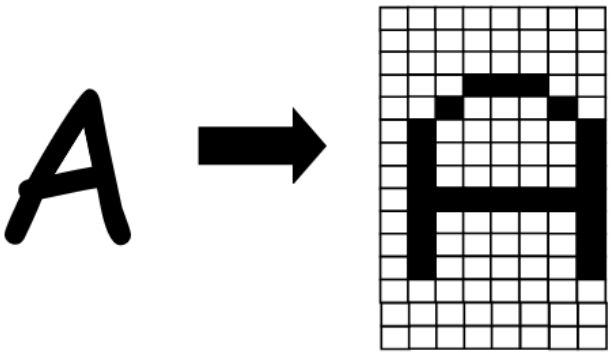

Figure 4: Bitmap representation by 8x15matrix or by 120 vectors with 0 and 1 coordinates.

Recently neural networks are used in OCR systems for classification. They are a network of several parallel interconnected adaptive neural processors. Its computational speed is higher compared to other techniques due to the parallel nature. The feature set enters the network at the input layer. Each element of the layer computes the weighted sum of its input and transforms it into an output by a nonlinear function. During the training process, the weights at each connection are adjusted until a desired output is obtained [10].

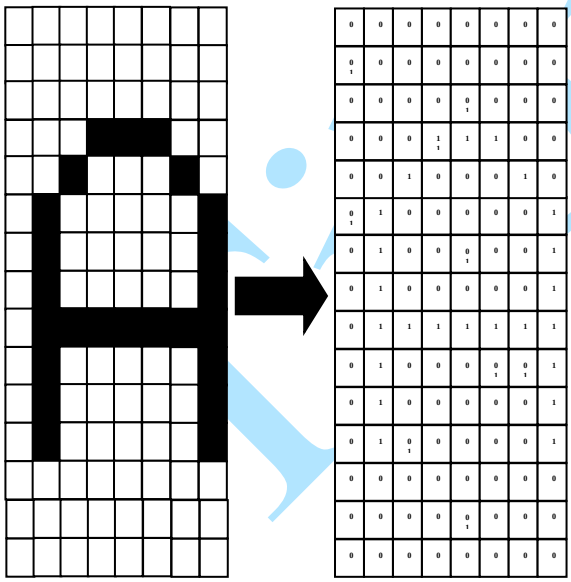

Figure 5: Binarization of Hausa Alphabet A Character

6. MATCHING ALGORITHM

Matching technique enables the creation of a template for all printed possible Hausa input characters. During recognition process, the printed input character is compared to each of the Hausa printed characters in the data set to get either the data set printed character with the closest match or an exact match of the input character [4]. If for example I(x, y) is the input printed character and Tn(x, y) is the data set character n, then the matching function s(I, Tn) should return a value indicating how correct the template n matches the input printed character. Some basic matching functions are based on the following formulas [13]:

$$S(I,T_n)=\sum_{i=0}^{w}\sum_{j=0}^{h}|I(i,j)-T_n(i,j)| \quad (1)$$

$$S(I,T_n)=\sum_{i=0}^{w}\sum_{j=0}^{h}(I(i,j)-(i,j))^2 \quad (2)$$

$$S(I,T_n)=\sum_{i=0}^{w}\sum_{j=0}^{h}I(i,j)-T_n(i,j) \quad (3)$$

$$S(I,T_n)=\frac{\sum_{i=0}^{w}\sum_{j=0}^{h}(I(i,j)-|ID(T_n(i,j)-|T_nD)}{\sqrt{\Sigma\ \Sigma\ (I(i,j)-|ID^2)}\sqrt{\Sigma\ \Sigma\ (T_n(i,j)-|T_nD^2}} \quad (4)$$

Matching formula: (1) City block, (2) Euclidean, (3) Cross Correlation, (4) Normalized Correlation.

Printed character recognition is always achieved by detecting which Tn gives the best value of matching function, s(I ,Tn) . The technique can only be executed successfully if the printed input character and the stored printed character templates are of the similar or same font. Matching of printed templates characters and printed input character can be done on binary printed characters, threshold printed characters or gray-level printed characters.

Template matching techniques were formed as a response to the issue of printed object recognition, and they contain implicitly the knowledge of similarity object comparison [12].

The basic idea behind the template matching method is the reference points. Reference points are points at the center of spatial regions in three-dimension space. For example, if the regions were defined to have x, y, and z center points, and three distance values, one for each of the three axes. However, by alternately subtracting and adding the distance values along the appropriate axis about the center point, a region in the shape of a cube is formed.

The template matching algorithm implementation takes the following steps:

a. The printed character image from the detected string is selected.
b. The printed character image is reduced to the size of the first template (the image is re-scaled).
c. The matching metric is calculated.
d. Then the highest match found is stored. If the printed character image does not match repeat step three above.
e. The index of the best match is stored as the recognized printed character image.

7. HAUSA LANGUAGE

Hausa language is one of the dominant languages in sub-Sahara Africa [1]. Hausa often used for human communication in the Northern part of Nigeria and other sub-Sahara Africa [1].

Table 1: Hausa Digits and its English Digits Equivalent.

Digits	Hausa Word of Digits	English Meaning of the Digits
0	Sifiri	Zero
1	Ɗaya	One
2	Biyu	Two
3	Uku	Three
4	Hudu	Four
5	Biyar	Five
6	Shida	Six
7	Bakwai	Seven
8	Takwas	Eight
9	Tara	Nine
10	Goma	Ten
11	Goma sha ɗaya	Eleven
20	Ashirin	Twenty
21	Ashirin da ɗaya	Twenty one
30	Talatin	Thirty
40	Arba'in	Fourty
50	Hamsin	Fifty
60	Sittin	Sixty
70	Saba'in	Seventy
80	Tamanin	Eighty
90	Tis'in	Ninety
100	Ɗari/Ɗari ɗaya	One Hundred
1000	Dubu/Dubu ɗaya	One Thousand
1000,000	Miliyan	One Million
1000,000,000,000	Biliyan	One Billion

In the same vein, Hausa language is spoken in countries like Congo, Ghana, Burkina-Faso, Cameroon, Niger, Benin, Sudan, Central African Republic, Togo and Chad [1]). Hausa is a tone based language which means that one word with the different tones could have different meanings according to pitch differences in syllables. Hausa tones are indicated in the written script with accent marks been placed on top of vowels [2]. Hausa consists of 35 character alphabets namely: (A/a, B/b, C/c, D/d, E/e, F/f, G/g, H/h, I/i, J/j, K/k, L/l, M/m, N/n, O/o, R/r, S/s, T/t, U/u, W/w, Y/y, Z/z) plus Ɓ/ɓ, Ɗ/ɗ, Ƙ/ƙ, 'Y/'y called hooked or gluttonised sounds, ['] called a glottal stop, every word written with an initial vowel in Hausa actually begins with a glottal stop, so strictly speaking, the word 'no' should be written 'a'a. There are seven basic digraphs in Hausa which are kw, sh, ky gy ƙy, ƙw, and ts. Hausa has five short and five long vowels which are: a, e, i, o, u; and aa, ee, ii, oo, uu. There are three basic tones in Hausa, namely: low tone, high tone and id/falling tone [1]. Additionally, Hausa distinguishes between short and long vowels which can also affect word meaning. Neither the vowel lengths nor the tones are marked in Hausa most conventional writings in text books and newspapers.

8. EXPERIMENTAL RESULTS

During training phase 25 printed alphabets representative documents each, were used. In order to evaluate the performance of the segmentation procedure was described in the study. The performance evaluation is based on counting the number of matches of printed alphabets detected by the algorithm with those in the ground truth [8].

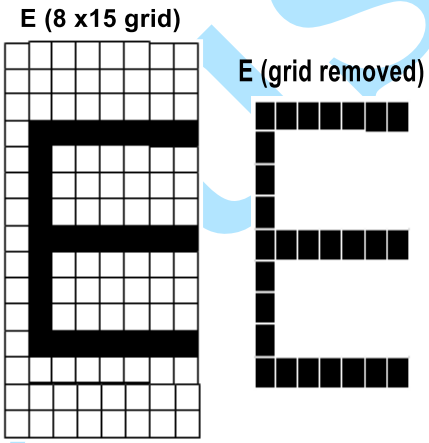

Figure 6: Alphabet E with grid and without grid

A match is considered only if the matching score is equal to or above the evaluator's acceptance threshold Tα. The performance is recorded in terms of detection rate (DR) and recognition accuracy (RA), while as an overall measure the Fmeasure (FM) which is a weighted harmonic mean of detection rate and recognition accuracy is used (Equation 6).

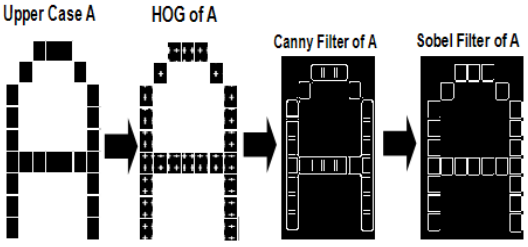

Figure 7: Alphabet A during Pre-processing

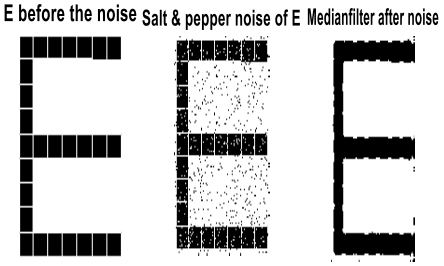

Figure 8: Alphabet E with and with Noise Removed

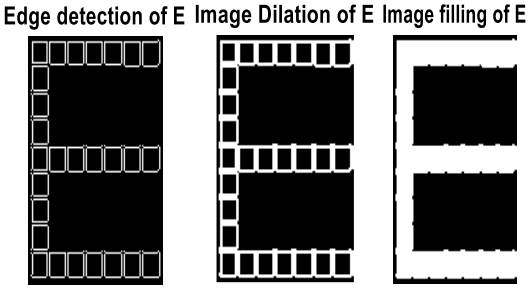

Figure 9: Alphabet E during Morphological Operations

A global performance metric SM is extracted by calculating the average values for FM metric for text line and word segmentation. Table 2 shows the results where the acceptance threshold is set to Tα =90. Table 3 depicts the results of training phase.

$$FM = \frac{2.DR.RA}{DR + RA} \qquad 6$$

Table 2: Result Presentation

Train characters	No of Test	Correct test	Error test	Percentage
Upper case (25)	40	36	4	90%
Lower case (25)	40	36	4	90%
High tone (5)	40	30	10	75%
Low tone (5)	40	30	10	75%
Digits	40	37	3	93%
Special symbols	40	25	15	60%

CONCLUSION

A simple and effective template matching method for identification of Hausa printed characters method was introduced in the study. For recognition process, the extracted character was compared to each template in the database to find the closest representation of the input character. The matching metric was computed using 2-D correlation coefficients approach to identify similar patterns between the test image and the database images. Experimental results show that the proposed method is efficient for identification Hausa printed characters.

REFERENCES

[1] **D. A. Burquest** - An Introduction to the Use of Aspect in Hausa Narrative, Language in context: Essays for Robert E. Longacre, Shin Ja J. Hwang and William R. Merrifield (eds.), 1992.

[2] **M. Elshafei** - Toward an Arabic Text-to-Speech System, The Arabian Journal for Science and Engineering, Vol. No. 16, Issue No. 4B, pp. 565-83, 1991.

[3] **J. D. Foley, A. V. Dam, S. K. Feiner, J. F. Hughes, and R. L. Phillips** - Introduction to Computer Graphics. Addison-Wesley, 1994.

[4] **B. Gatos, A. Antonacopoulos, N. Stamatopoulos** - Handwriting Segmentation Contest, 9th International Conference on Document Analysis and Recognition (ICDAR), Curitiba, Brazil, pp. 1284-1288, 2007.

[5] **R. Gonzalez and R. E. Woods** - Digital Image Processing. Addison-Wesley, second edition, 2002.

[6] **W. K. Pratt.** Digital Image Processing. John Wiley and Sons, second edition, 1991.

[7] **J. R. Prasad** - Image Normalization and Preprocessing for Gujarati, Character Recognition IJCSN International Journal of Computer Science and Network, Volume 3, Issue 5, October 2014 ISSN (Online): 2277-5420 www.IJCSN.org, pp: 334-340, 2014.

[8] **M. Rabbani and P. W. Jones** - Digital Image Compression Techniques. SPIE Optical Engineering Press, 1991.

[9] **J. P. Serra** - Image analysis and mathematical morphology. Academic Press, 1982

[10] **G. Smith** - Optical character Recognition. Machine Vision, CSIRO Manufacturing and Infrastructure Technology, Locked Bag 9, Preston 3072 Australia, 2007.

[11] **J. Sauvola and M. Pietikainen** - Adaptive Document Image Binarization, Pattern Recognition 33(2), pp. 225–236, 2000.

[12] **G. Vamvakas, B. Gatos, N. Stamatopoulos, and S. J. Perantonis** - A Complete Optical Character Recognition Methodology for Historical Documents Computational Intelligence Laboratory, Institute of Informatics and Telecommunications, National Center for Scientific Research "Demokritos", GR-153 10 Agia Paraskevi, Athens, Greece.pp 525-532, 2008.

[13] **T. Y. Zhang, and C. Y. Suen** - A fast parallel algorithm for thinning digital patterns, Communications of the ACM, vol.27 (3), pp.236–240, 1984.