THE ANNALS OF "DUNĂREA DE JOS" UNIVERSITY OF GALATI
FASCICLE III, 2014, VOL. 37, NO. 1, ISSN 2344-4738, ISSN-L 1221-454X
ELECTROTECHNICS, ELECTRONICS, AUTOMATIC CONTROL, INFORMATICS

NEW APPROACH OF FEATURE EXTRACTION METHOD BASED ON THE RAW FORM AND HIS SKELETON FOR GUJARATI HANDWRITTEN DIGITS USING NEURAL NETWORKS CLASSIFIER

Kamal Moro, Mohammed Fakir, Belaid Bouikhalene, Rachid El Yachi, Bader Dinne El Kessab

Information processing and telecommunication teams
Faculty of Science and Technology
Sultan Moulay Slimane University
523, Beni Mellal, Morocco

Abstract: This paper presents an optical character recognition (OCR) system for Gujarati handwritten digits. One may find so much of work for latin writing, arabic, chines, etc. but Gujarati is a language for which hardly any work is traceable especially for handwritten characters. Here in this work we have proposed a method of feature extraction based on the raw form of the character and his skeleton and we have shown the advantage of using this method over other approaches mentioned in this article.

Keywords: Optical character recognition, neural network, feature extraction, Gujarati handwritten digits, skeletonization**,** classification**.**

1. INTRODUCTION

The styles of writing numbers are highly different and they come in various sizes, shapes and fonts. The stability to identify these handwritten numbers in an automated or semi-automated manner has led to the development of an entirely different field of research known as the optical character recognition (OCR).

Very few attempts are found for the OCR activities related to Gujarati. It is a language from the Indo-Aryan family of languages, used by more than 50 million peoples in the Indian states of Gujarati. It is popular way of communication for Gujarati people, staying in all most every country of the world. A formal grammar of the precursor of this language was written by Jain monk and eminent scholar Hemachandracharya. It has a rich oral culture and a literary tradition which dates back to the tenth century. The construction of Gujarati can be considered somewhere between those of Hindi and Marathi. Fig.1 shows an example of Gujarati characters.

For developing a system to identify Gujarati handwritten digits, we have collected numerals 0-9 written in Gujarati scripts from a large number of writers. These numbers were scanned in 300 dpi by flatbet scanner. Initiallythey are in separate boxes of 50*30 pixels each. Since our problem is to identify

This paper was recommended for publication by Viorel Nicolau

handwritten digits, the first thing required is to bring all the characters in a standard normal form. This is needed because when a writer writes he may use different types of pen, papers, they may follow even different styles of writing etc.

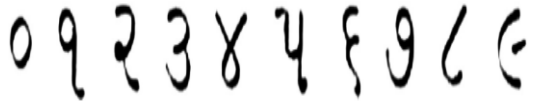

Fig.1. Gujarati digits à-9

The character recognition systems offer potential benefits by providing an interface that facilitates interaction between man and machine, these systems are based on algorithms essentially consist of three main steps: preprocessing, feature extraction and classification. Skeletonization is considered an essential part of the preprocessing but not essential, more research has used the skeletonization (hanmandly.M *et al*, 2007), (Moro.K *et al*, 2009), and others have chosen to deal directly on the raw form (Amrouch.M *et al*, 2009), (El Kessab.B *el al*, 2009), (Desai.A *et al*, 2010), (El yachi.R *et al*, 2010). We present the following method Gujarati character recognition, the feature extraction method used in this algorithm is applied to the raw form and on the skeleton, the average of these two vectors is the one that is used later in the classification step. Fig.2 summarizes the recognition process used.

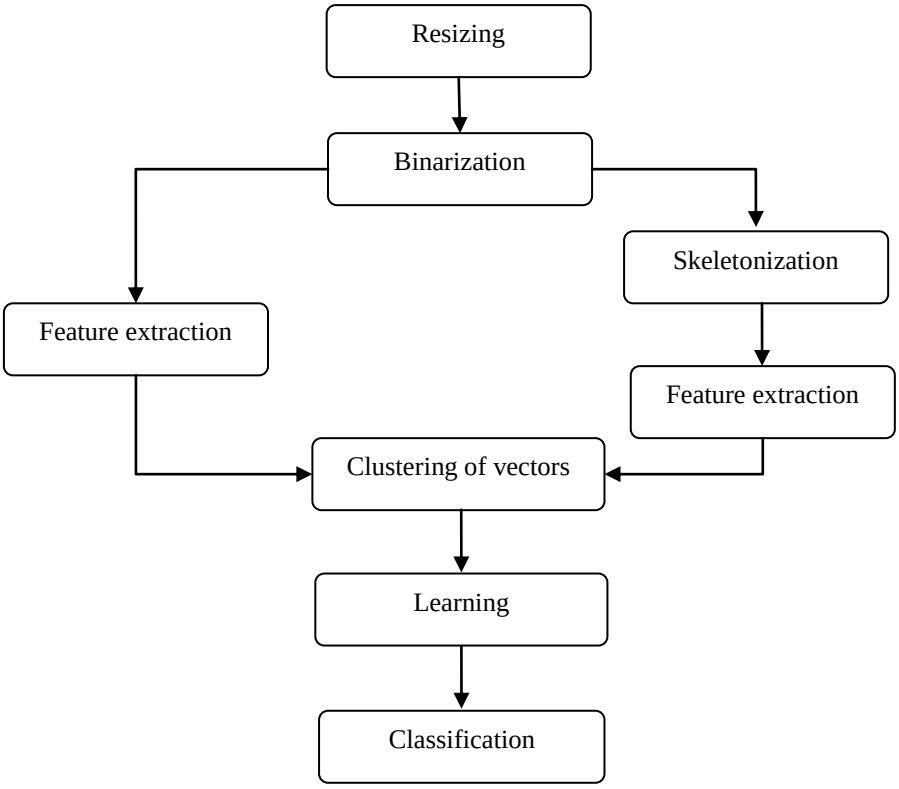

Fig.2. Recognition process

2. BINARIZATION

Binarization is an important step in any process of image analysis. A large number of binarization techniques have been proposed in the literature (Kefali.A *et al*, 2009), whose each of them is appropriate for a particular type of image. In this document, we use the method referred in (Wolf.C *et al*, 2002) which is to calculate the threshold of each pixel locally by following the formula:

$$(1)\ T = (1-k) * m + k * m + k * \frac{\sigma}{R(m-M)}$$

As k is set to 0.5, σ the standard deviation and m the average of all the pixels in the window, M is the minimum image grey level and R is the maximum deviation of grayscale on all Windows.

Fig.3 shows the result of the use of the Wolf algorithm (Wolf.C *et al*, 2002).

Fig.3. Binarization of a digit

3. SKELETONIZATION

Skeletonization is a class of algorithms used in shape analysis. It consists to reduce a form of a set of curves, called skeletons, centered in the original form. Skeletonization is an analysis of non-scalar form tool, which preserves the topological properties of the original shape and geometrical properties, depending on the method used.

In this document, we chose to use the Guo_Hall algorithm (Guo.Z *et al*, 1989), it uses the parallel approach of thinning, It preserves the topology and geometry.

X4	X3	X2
X5	P	X1
X6	X7	X8

Fig.4. A point P and its neighbords

By considering a point P (Fig.4) and its noted neighborhoods X1, X2, X3, X4, X5, X6, X7 and X8, the GUO_HALL algorithm is to remove parallel points of the object P checking the following conditions:

- P is 4-adjacent to the supplementary object
- (2) $2 \leq N(P) \leq 3$
- (3) $(x_2 \vee x_3 \vee \overline{x_8}) \wedge x_1 = 0$
- (4) $(x_6 \vee x_7 \vee \overline{x_4}) \wedge x_5 = 0$

with

$$(5)\mathrm{N}_1(\mathrm{P}) = (x_1 \vee x_2) + (x_3 \vee x_4) + (x_5 \vee x_6) + (x_7 \vee x_8)$$

$$(6)\mathrm{N}_2(\mathrm{P}) = (x_2 \vee x_3) + (x_4 \vee x_5) + (x_6 \vee x_7) + (x_8 \vee x_1)$$

$$(7) N(P) = Min\big(N_1(P), N_2(P)\big)$$

Fig.5 shows the result of the Guo Hall algorithm that is used in this document.

4. FEATURE EXTRACTION

In pattern recognition and in image processing, feature extraction is a special form of dimensionality reduction. When the input data to an algorithm is too large to be processed and it is

Fig.5. Skeletonization of a digits

suspected to be notoriously redundant (e.g. the same measurement in both feet and meters) then the input data will be transformed into a reduced representation set of features (also named features vector). Transforming the input data into the set of features is called feature extraction. If the features extracted are carefully chosen it is expected that the features set will extract the relevant information from the input data in order to perform the desired task using this reduced representation instead of the full size input.

For extracting features, we use the Box-approach in Refs (Hanmandly.M *et al*, 2007), (Hanmandlu.M *et al*, 2003), (Hnamandu.M *et al*, 2005). This approach requires the special division of the character image. The major advantage of this approach stems from its robustness to small variations, ease of implementation and relatively high recognition rate.

Each character image is divided into 24 boxes so that the portions of a numeral will be in some of these boxes. There could be boxes that are empty, as shown in Fig.6. English numeral 3 is enclosed in the 6*4 grid. However, all boxes are considered for analysis in a sequential order. By considering the bottom left corner as the absolute origin (0,0), the coordinate distance (vector distance) for the *k*th pixel in the bth box at location (i,j) is computed as

$$(8)\ d_{kb} = (i^2 + j^2)^{\frac{1}{2}}$$

By dividing the sum of distances of all black pixels present in a box with the total number of pixels in that box, a normalized vector distance (γ_b) for each box is obtained as

$$(9)\ \gamma_b = \frac{1}{n_b}\sum_{1}^{n_b} d_k^b \ , \mathrm{b} = 1 \ldots 24$$

where n_b is the total number of pixels in *b*th box. These vector distances constitute a set of features based on distances. Therefore, 24 y_b 's corresponding to 24 boxes will constitute a feature set. However, for empty boxes, the value will be zero.

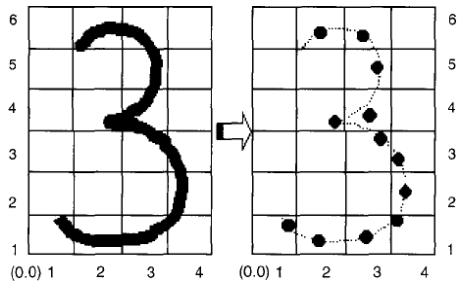

Fig.6. Portions of the numeral lie within some boxes while others are empty

A profile vector for a digit is shown here:

[0 0 0.6672 0.7017 0 0 1.9429 2.8865 0 0 0.5770 2.0683 0 0 0 3.3593 0 0 1.1157 0 2.4256 1.8303 0 2.4254 3.5686 0 4.0410 6.6789 9.9418 0 0 0 0 7.1634 0 0 0 0 8.0269 0 0 9.6223 0 8.6789 0 0 0 10.4075 3.5689 0].

The feature extraction method is applied to the raw form and on the skeleton, the average of these two vectors is the one that is used later in the classification step.

5. NEURAL NETWORKS

As (Lu Tan.C et al, 2004), (Sukhswam.M.B et al, 1995) have used neural network for character classification, in this work, neural network is suggested. A feed forward back propagation neural network is used for Gujarati numeral classification. This proposed multi-layered neural network consists of three layers with 50, 30, 10 for distance method and 118, 60 respectively, and 10 neurons for sum method. The input layer is the layer which accepts the profile vector which is of 1*50 and 1*118 in size. As this network is used for classification of 10 digits, it has 10 neurons in the output layer, the function sigmoid as function of activation at the step of the layer entry and hidden, with $\alpha = 0.1$,

$$(10)\ f(x) = \frac{1}{1 + e^{-\gamma x}}$$

logsig at the step of the output layer and we fixed the constant learning to $\gamma = 0{,}1$.

For this experiment, a total of 300 responses were taken into consideration. For training, the features are abstracted first for all of these images of digits.

To prevent the over learning, a set of validation characters is used. These characters have to define on the algorithm the best values of synaptic weights. Data validations are neutral in the determination of the weight; they serve only to stop a previous iteration, before the start of the over learning. In our case, we used 100 characters of validation.

6. EXPERIMENTAL RESULTS

We present in this part the experimental results of the recognition process in Fig.2, and the comparison with other conventional approaches, namely the treatment with using skeleton and without skeleton. Table 1 summarizes these results.

Table 1 Result Summary

Sets	**Without Skeleton**	**With Skeleton**	**Proposed Method**
600	85,33%	80,5%	85,67%

Tables 2, table 3 and table 4 give more details on the recognition rate of each class of characters and table 5 shows different confusion between characters. It is considered that a character is confused with another if the error rate exceeds 10%.

7. CONCLUSION

In this work we feed forward back propagation neural network is proposed for the classification of the Gujarati numerals. Various techniques are used in the preprocessing step before implementing classification of numerals. The overall performance of this proposed network is as high as 85,33% with proposed method, 80,5% with skeleton and 85,33% without skeleton.

The performance of each method of classification is based on the extraction of characteristics. In our perspective, we intend to apply other techniques of extraction in the recognition process and use hidden Markov networks and Bayesian Networks at the level of the classification.

Table 2 Performance of the process without using skeletonization

Numbers	Without Skeleton										
	0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**Success(%)**
0	54	0	0	0	0	0	0	4	2	0	90,00
1	0	54	2	0	1	3	0	0	0	0	90,00
2	0	11	41	0	5	2	0	0	1	0	68,33
3	0	0	0	55	0	0	2	3	0	0	91,67
4	0	1	0	0	48	2	0	2	7	0	80,00
5	0	10	6	0	0	44	0	0	0	0	73,33
6	9	0	0	3	0	0	45	1	1	1	75,00
7	3	0	0	2	0	0	0	55	0	0	91,67
8	0	0	0	0	2	0	0	0	52	6	86,67
9	0	0	0	0	3	0	0	0	26	31	51,67

Table 3 Performance of the process with using skeletonization

Numbers	With Skeleton										
	0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**Success(%)**
0	54	0	0	1	0	0	1	0	3	1	90,00
1	0	53	1	0	2	4	0	0	0	0	88,33
2	0	11	44	1	4	0	0	0	0	0	73,33
3	0	0	0	57	0	0	0	3	0	0	95,00
4	0	3	0	3	47	0	0	0	5	2	78,33
5	0	11	6	1	0	41	1	0	0	0	68,33
6	0	0	0	9	0	2	46	3	0	0	76,67
7	2	0	0	1	1	0	1	54	0	1	90,00
8	0	0	0	0	2	0	0	0	56	2	93,33
9	0	1	0	1	0	0	0	0	14	44	73,33

Table 4 Performance of the process with proposed method

Numbers	Proposed Method										
	0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**Success(%)**
0	52	0	0	0	0	0	0	4	3	1	86,67
1	0	53	2	0	2	3	0	0	0	0	88,33
2	0	8	43	0	5	2	0	0	1	1	71,67
3	0	0	0	55	0	0	1	4	0	0	91,67
4	0	1	0	0	49	0	0	1	3	6	81,67
5	0	11	5	0	0	42	2	0	0	0	70,00
6	1	0	0	5	0	0	48	0	0	6	80,00
7	1	0	0	2	0	0	0	57	0	0	95,00
8	0	0	0	0	1	0	0	0	58	1	96,67
9	0	0	0	0	2	0	0	0	17	41	68,33

Table 5 Confusion of characters

Digits	Without Skeleton	With Skeleton	Proposed Method
0	Any	Any	Any
1	Any	Any	Any
2	1	1	1
3	Any	Any	Any
4	8	Any	9
5	1,2	1,2	1
6	0	3	9
7	Any	Any	Any
8	9	Any	Any
9	8	8	8

8. REFERENCES

Hanmandly.M, Ramana Murthy.O.V (2007). *Fuzzy model based recognition of handwrittern numerals*, Pattern recognition 40, pp: 1840-1854

Moro.K, Fakir.M, Bouikhalene.B, Safi.S, (2009). *Skeletonization Methods Evaluation for the Recognition of Printed Tifinaghe Characters*, Sitacam '09 Agadir Morocco, pp: 33-47

Amrouch.M, Es Saady.Y, Rachidi.A, El Yassa.M, Mammass.D.(2009), *Apprentissage Markovien et Neuronal: cas des caractères amazighes imprimés*, Sitacam'09 Agadir Morocco, pp: 58-67.

El Kessab.B, Bouikhalene.B, Fakir.M, Safi.S. (2009) *Reconnaissance des caractéres Tifinaghe par l'utilisation des réseaux de neurones multicouches*, Sitacam'09 Agadir Morocco, pp: 68-83.

Desai.A. (2010). Gujarati *handwritten numeral optical character reorganization through neural network*, Pattern Recognition 43, pp: 2582-2589.

El Yachi.R, Moro.K, Fakir.M, Bouikhalene.B. (2010). *On the recognition of Tifinaghe Scripts.* Journal of Theoretical and Applied Information Technology, pp : 61-66

Kefali.A, Sari.T, Sellami.M. (2009). *Evaluation de plusieurs techniques de seuillage d'images de documents arabes anciens*, 5ieme Symposium Internationnal-IMAGE'2009

Wolf.C, Jolion.J.M, Chassaing.F. (2002). *Extraction de texte dans des vidéos : le cas de la binarisation*, 13ème Congrès Francophone de Reconnaissance des Formes et Intelligence Artificielle, pp : 145-152.

Guo.Z, Hall.R.W. (1989). *Parallel thinning with two subiteration algorithms*. *Comm.* ACM, 32(3), pp: 359–373.

Hanmandlu.M, Murali Mohan.K.R, Chakraborty.S, Goyal.S, Roy Choudhury.D. (2003). *Unconstrained handwritten character recognition based on fuzzy logic*, Pattern Recognition 36 (3), pp: 603-623.

Hanmandlu.M, Yusof.M.H.M, Vamsi Krishna.M. (2005). *Off-line signature verification and forgery detection using fuzzy modelling*, Pattern Recognition 38 (3), pp: 341-356.

Luh Tan.C, Juntan.A. (2004). *Digit recognition using neural networks*, Malaysian Journal of Computer Science, 17 (2), pp: 40–54.

Sukhswami.M.B, Seetharamulu.P, Pujari.A. (1995). *Recognition of Telugu characters using neural networks*, International Journal of Neural Systems, 6 (3), pp: 317–357.