GUJARATI HANDWRITTEN NUMERAL OPTICAL CHARACTER THROUGH NEURAL NETWORK AND SKELETONIZATION

Kamal MORO*, Mohammed FAKIR, Badr Dine EL KESSAB, Belaid BOUIKHALENE, Cherki DAOUI

(dont delete this line. It is used to insert authors detail)

Abstract— ***This paper deals with an optical character recognition (OCR) system for handwritten Gujarati numbers. One may find so much of work for Indian languages like Hindi, Kannada, Tamil, Bangala, Malayalam, Gurumukhi etc, but Gujarati is a language for which hardly any work is traceable especially for handwritten characters. The features of Gujarati digits are abstracted by four different profiles of digits. Skeletonization and binarization are also done for preprocessing of handwritten numerals before their classification. This work has achieved approximately 80,5% of success rate for Gujarati handwritten digit identification.***

***Index Terms*—–Optical character recognition, neural network, feature extraction, Gujarati handwritten digits, skeletonization, classification.**

I. INTRODUCTION

Gujarati belonging to Devnagari family of languages, which originated and flourished in Gujarat a western state of India, is spoken by over 50 million people of the state. Though it has inherited rich cultural and literature, and is a very widely spoken language, hardly any significant work has been done for the identification of Gujarati optical characters. The Gujarati script differs from those of many other Indian languages not having any shirolekha (headlines). Gujarati numerals do not carry shirolekha and it applies to almost all Indian languages. The numerals in Indian languages are based on sharp curves and hardly any straight lines are used. Fig.1 is a set of Gujarati numerals.

As it is visible in Fig.1, Gujarati digits are very peculiar by nature. Only two Gujarati digits one(1) and five(5) are having straight line, making Gujarati digit identification a little more difficult. Also Gujarati digits often invite misclassification. These confusing sets of digits areas shown in Fig. 2.

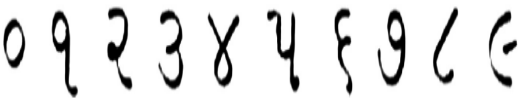

Fig. 1 Gujarati digits 0-9

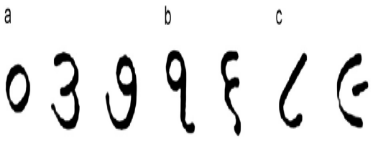

Fig. 2 Confusing Gujarati digits

This paper addresses the problem of handwritten Gujarati numeral recognition. Gujarati numeral recognition requires binarization and skeletonozation as preprocess. Further, profiles are used for feature extraction and artificial neural network (ANN) is suggested for the classification.

II. DATABASE

For handwritten English numerals, we have the CEDAR (Centre of Excellence for Document Analysis and Recognition at the University of New York at Buffalo, USA) numeral database. It contains approximately 5000 samples of numerals. It contains approximately 5000 samples of numerals. The samples are originally collected from US postal ZIP codes found on letters. As there is no standard database available at the moment for Gujarati.

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 a flatbed scanner. Initially they are in separate boxes of 50*30 pixels each. Since our problem is to identify 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 pens, papers, they may follow even different styles of writing etc.

III. BINARIZATION

Binarization is often the important first step in any process of character recognition. A large number of binarization techniques have been proposed in the literature [1], which each is appropriate to a particular type of images. It has as a goal to reduce the amount of information present in the image, and keep only the relevant information.

According to several research work [2,3], the techniques of binarization of grayscale images can be classified into two categories: overall threshold, where a single threshold is used in the entire image to the devise in two classes (text and background), and local threshold where the values of the thresholds are determined locally, pixel-by-pixel or well region by region. In this document,

Information processing and telecommunication teams, Faculty of Science and Technology, University of Sultan Moulay Slimane, Beni Mellal, Morocco
kamalmoro@hotmail.com, fakfad@yahoo.fr, bbouikhalene@yahoo.fr, bade10@hotmail.fr, daouic@hotmail.fr

we use the method referred in [4] which is to calculate the threshold of each pixel locally by following the formula:

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

As k is set to 0.5, the difference type 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.

IV. SKELETONIZATION

A fundamental problem in pattern recognition is a synthetic representation of this. In many cases, to work on the raw form is laborious and unnecessary. It is much more advantageous in terms of time and quality to work with a refined shape. The notion of skeleton was introduced for this effect. In the ongoing plan, the skeleton of a shape is a set of lines passing through the middle. This is the concept of median axis of a continuous form introduced by Blum [5].

There are currently a wide variety of methods to construct skeletons from shapes, which topological thinning who is to remove the points of the outline of the shape, while preserving its topological characteristics. In this document, we chose to use the Guo_Hall algorithm [6], it uses the parallel approach of thinning, It preserves the topology and geometry, it is cited in [7].

X4	X3	X2
X5	P	X1
X6	X7	X8

Fig.3 A point P and its neighborhoods

A point P (Figure 3) 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
- $2 \leq N(P) \leq 3 \quad (2)$
- $x_2 \vee x_3 \vee \overline{x_8}) \wedge x_1 \quad (3)$
- $x_6 \vee x_7 \vee \overline{x_4}) \wedge x_5 \quad (4)$

with

$$N_1(P) = (x_1 \vee x_2) + (x_3 \vee x_4) + (x_5 \vee x_6) + (x_7 \vee x_8) \quad (5)$$

$$N_2(P) = (x_2 \vee x_3) + (x_4 \vee x_5) + (x_6 \vee x_7) + (x_8 \vee x_1) \quad (6)$$

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

V. FEATURE EXTRACTION

Feature extraction is the most important phase in the field of the recognition of characters, [8] have used invariant moments for recognition of Tifinagh characters, [9] have used vectors of cavity and have also applied on the Tifinagh characters. In this document, we have chosen to use a method that is both simple and effective, this method consist in doing the sum of the values of pixels at level horizontal, vertical and two diagonal. The cavities show the way to summon the pixels to an image 3 x 3. For example, in considering the form of the Fig .5, the Table 1 shows vector extraction following the Fig .4 patterns.

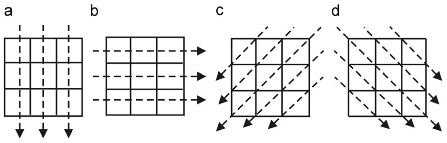

Fig. 4 Pattern profile of 3x3 pattern matrix

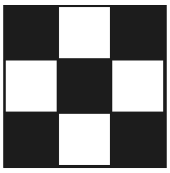

Fig. 5 3x3 Pattern

Figure 3	Extraction vector
A	(2,1,2)
B	(2,1,2)
C	(1,0,3,2,1)
D	(1,0,3,2,1)

Table 1 Extraction vectors

In our knowledge, only [10] has used this feature extraction method on the Gujarati characters and applied it on the raw form of the character instead of his skeleton. In this work, For Gujarati numeral recognition profile vector is created for all the digits which are converted into 16*16 pixels after preprocessing.

VI. NEURAL NETWORKS

As [11, 12] 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 118, 60, and 10 neurons, respectively. The input layer is the layer which accepts the profile vector which is of 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 $\alpha = 0.1$,

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

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

VII. TRANING OF NETWORKS

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. A profile vector for a digit five is shown here:
[0 0 0 0 0 5 2 2 1 1 1 1 1
1 1 1 1 1 1 1 0 0 0 0 2 3 2
2 1 1 1 1 1 1 1 1 1 1 1 1 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 7 3 3 2 3 1 1 0 1 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 3 1 1 1 1 0 1 0 1 0 1
0 1 0 1 0 1 1 0 1 1 0 1 1
1 0 1 1 0 0 0 0].

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.

In the Fig. 6 the complete process of Gujarati numeral optical character recognition is shown.

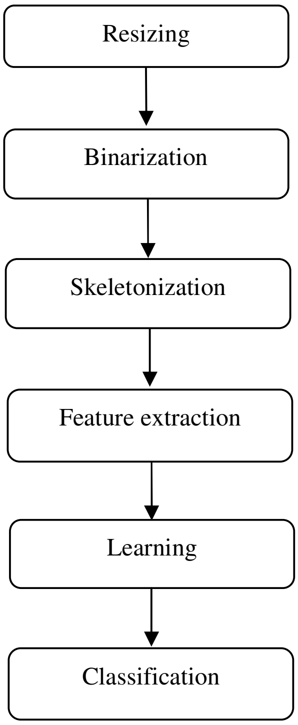

Fig. 6 Recognition process

VIII. EXPERIMENTAL RESULTS

As mentioned above this network was trained for total 30 sets of digits, and was tested for 60 other new sets of digits. In total the network was trained by 300 digits and tested for 600 digits.

Initially, we apply the binarization on the character, this operation aims to eliminate the various intensities of gray pixels of the image to make binary, Fig.7 shows the result of the use of the Wolf algorithm [4].

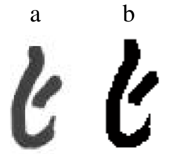

Fig. 7 Binarization of a digit, a: bifor binarization, b: after binarization

After binarization, we begin the skeletonization step, this approach is designed to present the form with a minimum of informations; the Fig.8 shows the result of the Guo_Hall [6] algorithm that is used in this document.

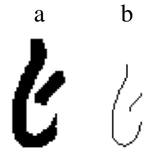

Fig. 8 Skeleton of a digit, a: bifor sketetonization, b: after skeletonization

The success rate of the proposed network is 80,33%, the results are summarized in table 2.

Sets	Nos of digits	Type of sets	Success rate (%)
30 sets of digits	300	Training sets	100
60 sets of digits	600	Testing sets	80,33

Table 2 Experimental results

Let us examine the results obtained for the different digits. Table 3and table 4 shows success rate for testing handwritten digits.

Digits	0	1	2	3	4
0	44	0	0	0	0
1	0	51	3	4	0
2	0	9	42	0	2
3	0	1	0	50	0
4	0	0	0	2	54
5	2	5	1	1	2
6	0	0	2	6	0
7	0	0	0	9	1
8	0	0	0	0	0
9	0	0	0	0	9

Table 3 Result summary

5	6	7	8	9	Success(%)
0	0	5	0	11	73,33
2	0	0	0	0	85,00
1	5	0	1	0	70,00
0	9	0	0	0	83,33
1	0	2	0	1	90,00
48	1	0	0	0	80,00
0	48	0	0	4	80,00
0	0	49	0	1	81,67
0	0	0	60	0	100
0	1	6	8	36	60,00

Table 4 Result summary (suite)

The first note that there is that digits 9 are confused with the 8 and 4, also, digits 1 and digits 4 are slightly confused with other characters. Table 5 shows the different confusion of characters, It is considered that a character is confused with who is treated if the error rate exceeds 10%.

Character treated	Character confused
0	9
1	Any
2	1
3	6
4	Any
5	Any
6	3
7	3
8	Any
9	4,8

Table 5 Confusion of characters

IX. 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 80,5%.

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.

REFERENCES

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

[2] Arica, N., Yarman-Vural, F. T., *An overview of character recognition focused on off-line handwriting*, IEEE transactions on systems, man and cybernetics - part C: Applications and reviews, 31(2), 2001, pp: 216-233

[3] Khurshid, K., Siddiqi, I., Faure, C., Vincent, N., *Comparison of Niblack inspired Binarization methods for ancient documents*, In 16th International conference on Document Recognition and Retrieval, USA, 2009

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

[5] BLUM, H., *A transformation for extracting new descriptions of shape*, In Models for the Perception of Speech and Visual Form, MIT Press, 1967, pp: 362–380

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

[7] Moro, K., Fakir, M., Bouikhalene, B., Safi, S., *Skeletonization Methods Evaluation for the Recognition of Printed Tifinaghe Characters*, In Sitacam '09 Agadir Morocco, 12-13 December 2009, pp: 33-47

[8] El Yachi, R., Moro, K., Fakir, M., Bouikhalene, B., *Utilisation des moments invariants et la programmation dynamique pour la reconnaissance des caractéres Tifinagh*, In Journal of Theoretical and Applied Information Technology, 2010, pp : 61-66

[9] El Kessab, B., Daoui, C., Moro, K., Bouikhalene, B., Fakir, M., *Recognition of Handwritten Tifinaghe Characters using a Multilayer Neural Networks and Hidden Markov Model*, Global Journals (USA), pp: 24-31, 2011

[10] Desai, A., *Gujarati handwritten numeral optical character reorganization through neural network,* In Pattern Recognition 43-2010, pp: 2582-2589

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

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

Kamal MORO has obtained the degree of Master in business intelligence in 2009 in Sultan Moulay Slimane University, Faculty of Sciences and Techniques of Beni Mellal. Currently, doctoral student at the FST of Beni Mellal, Morocco, its research focus on the pattern recognition and artificial intelligence.

Mohammed FAKIR obtained the degree of Master in electrical engineering from the University of technology of Nagaoka in 1991 and the degree Ph.d. in electrical engineering from the University of Cadi Ayyad. He is team Hitachi Ltd., Japan between 1991 and 1994. Currently, he is a professor at the Faculty of Sciences and technology, University Sultan Moulay Slimane, Morocco. His research concern the recognition and artificial intelligence.

Bader Dine El Kessab has obtained the degree of Master in business intelligence in 2009 in Sultan Moulay Slimane University, Faculty of Sciences and Techniques of Beni Mellal. Currently, doctoral student at the FST of Beni Mellal, Morocco, its research focus on the pattern recognition and artificial intelligence.

Belaid BOUIKHALENE obtained the PhD degree in mathematics in 2001 and degree of Master in Science of Informatics in 2005 on the University of Ibn Tofel. Currently, he is a professor at the University Sultan Moulay Slimane, Morocco, his research concerning the pattern recognition and artificial intelligence.

Cherki DAOUI obtained the PhD degree in mathematics in 2002 at the Mohamed V University. Currently he is a professor at the University Sultan Moulay Slimane, Morocco. His research interests include mathematics, the desired operational and pattern recognition.

Daftar Penulis

PEDOMAN PENULISAN ARTIKEL / MAKALAH
JURNAL SISTEM KOMPUTER

1. Redaksi menerima tulisan/ naskah karya ilmiah bidang rumpun ilmu komputer dari kalangan staf pengajar Fakultas Teknik Universitas Diponegoro dan dari kalangan umum.
2. Jurnal Sistem Komputer dapat menerima naskah-naskah karya ilmiah yang berupa:
   - Hasil Penelitian yang asli
   - Catatan Penelitian
   - Kajian Pustaka yang mempunyai kontribusi yang baru bagi ilmu pengetahuan
   - Komentar/ kritik tentang naskah yang pernah dimuat oleh Jurnal Sistem Komputer
3. Naskah yang dikirim ke Redaksi Jurnal Sistem Komputer akan di- *review* terlebih dahulu oleh Dewan Redaksi atau Mitra Bestari atau Pakar-Pakar di bidangnya. Keputusan diterima atau tidak diterimanya suatu artikel merupakan hak dari Dewan Redaksi berdasarkan saran-saran dari Reviewer.
4. Proses review akan dilaksanakan oleh Dewan Redaksi sehingga untuk kelancaran transfer file sebaiknya lewat e-mail agar lebih cepat prosesnya dan korespondensi akan ditujukan kepada alamat penulis pertama atau *Correspending Author* (setiap makalah harus ditandai siapa yang menjadi Penulis Penanggungjawabnya). Penulis harus segera memperbaiki artikel sesuai petunjuk Referees dan petunjuk penulisan jurnal dan dikirimkan kembali dengan segera.
5. Makalah yang ditulis harus sesuai format yang ditentukan (mengikuti standard Transaction Journal IEEE) dan harus mengandung komponen-komponen berikut (sesuai urutan):
   - Judul, Nama Penulis, Kata Kunci, Abstrak (dalam bahasa Inggris yang baik dan benar)
   - Pendahuluan
   - Bahan dan Metodelogi Penelitian
   - Hasil dan Pembahasan
   - Kesimpulan
   - Ucapan terima kasih (jika ada)
   - Daftar Pustaka
   - Biografi singkat penulis di akhir bagian
6. Naskah dapat ditulis dalam Bahasa Indonesia atau Bahasa Inggris. Naskah berisi maksimum 10 halaman kuarto (A4) termasuk gambar dan tabel, dikirimkan sebanyak dua eksemplar disertai dengan rekaman dalam disket ukuran 3.5" atau dalam CD. Naskah yang dikirimkan harus sudah siap untuk dicetak (*Camera ready*).
7. Artikel harus ditulis pada kertas ukuran HVS ukuran A4 (210 x 297 mm) dan dengan format margin kiri 25 mm, margin kanan 20 mm, margin bawah 30 mm dan margin atas 20 mm, serta harus diketik dengan jenis huruf Times New Roman dengan font 10 pt (kecuali judul), satu spasi dan dalam format dua kolom (kecuali judul, nama penulis, abstrak dan kata kunci dalam format satu kolom) yang terpisah sejauh 10 mm.
8. Judul tulisan dibuat sesingkat mungkin dan jelas, menunjukkan dengan tepat masalah yang hendak dikemukakan, tidak memberi peluang penafsiran yang beraneka raga.
9. Nama Penulis ditulis dibawah Judul Artikel tanpa disertai gelar akademik. Apabila Penulis lebih dari satu orang, nama-nama ditulis pada satu baris dipisahkan oleh koma. Nama instansi ditulis di catatan kaki halaman pertama makalah.
10. Abstrak (dalam bahasa Inggris yang baik dan benar) harus memuat inti permasalahan yang dikemukakan, metode pemecahannya, dan hasil-hasil yang diperoleh serta kesimpulan, dan tidak lebih dari 200 kata.
11. Kata-kata atau istilah asing yang digunakan huruf miring (Italic). Paragraf baru dimulai pada ketikan ke enam dari batas kiri, sedangkan antar paragraf tidak diberi antara. Semua bilangan ditulis dengan angka, kecuali pada awal kalimat. Tabel dan gambar harus diberi keterangan yang jelas.

9 772087 468004