Optical Recognition of Isolated Machine Printed Sindhi Characters using Fourier Descriptors

Nasreen Nizamani[1], Mujtaba Shaikh[2], Jawed Unar[3], Ehsan Ali[4], Ghulam Mustafa Bhutto[5], Abdul Rafay[6]
Department of Electronic Engineering, QUEST, Nawabshah, Pakistan[1, 2, 4, 5, 6]
Information Technology, QUEST, Nawabshah, Pakistan[3]

***Abstract*—The scale invariance characteristics play an essential role in pattern recognition applications, for example in computer vision, OCR (Optical Character Recognition), electronic publication, etc. In this paper, the shape based feature extraction techniques are used in terms of invariant properties and the region based FD (Fourier Descriptors) have been used for the recognition of isolated printed Sindhi characters. There are 56 isolated characters in Sindhi language than can be categorized to 20 different classes considering the shape of the base of each character. In this work, the dataset contains 4704 images of isolated printed Sindhi characters. The simulation result shows that the proposed method is a capable discriminating algorithm of Similar Sindhi characters and can easily extract the scale invariant features.**

Keywords—Features extraction; Sindhi optical character recognition; Fourier Descriptors; machine printed Sindhi characters

I. INTRODUCTION

The Optical character recognition (OCR) is a potential field of computer vision and pattern recognition. The OCR is an electronic conversion of the photographed or scanned images of printed or typewritten text into computer-readable text. The OCR can develop the interface between human and machines in several applications; for example, the office automation, business, cheque verification and data entry [1]. Usually, a typical OCR system comprises of different modules that carry out the recognition procedure. These modules are: *(i) Image acquisition module* – reads the gray scale or color images, *(ii) pre -processing* – this module carries out different operations (binarization and segmentation) operations on the input image, *(iii) Feature extraction* – this module extracts the complicated features from the input image, *(iv) Recognition / classification:* this module performs the recognition / classification task, and *(v) post-processing.* In OCR field, the feature extraction for characters plays an essential role in system. Different methods for recognition of different languages for instance Latin, Chinese and Arabic documents have been proposed in [2,3]. However, the recognition of Sindhi character is still demanding. The main challenges in Sindhi text are Perso-Arabic script having more characters dots and variation of placement and orientation of dots, four dotted characters, a large set of character recognition, the same base shape with variation in number, placement and orientation of dots [4]. In Sindhi text the individual Sindhi characters are rarely used however without individual characters the most of the times sentence has not a complete sense/meaning. Sindhi isolated characters like ﷽ and ﷺ, however, without these isolated characters the sentence is incomplete. The Sindhi alphabets have 52 common characters as shown in Fig. 1**.** In the Arabic script, the base shape "ب" is used only for three characters, "ب", "ت", and "ث" whereas, in the Sindhi script the same base shape is used for 9 characters, ب، ٻ، پ، ت، ٽ، ٹ، ٿ، ٺ، پ. Thus, an Arabic OCR may be able to recognize only three characters for this particular base shape, as well as their changing character forms, Whereas, a Sindhi OCR needs to be able of recognize nine different characters for the same base shape as given in Table I. The few characters like "ت" and "ٽ", "ٺ" and "ٹ" are more complex because they have the same base shape, only the difference to recognize is the dots orientation in each character. The similar problem is with other shape like"چ"and "ڇ".

The rotation, size scale (RST) invariant feature extraction methods are described [5, 6]. The Projection transformation process is more proficient because few regular arithmetic operations are performed, and this method is called transformation ring projection (TRP) method [5], and this approach transforms two dimentional (2-D) patterns into one-dimentional (1-D) patterns. This approach has a low recognition rate for similar characters [14]. For that reason, it is difficult to develop an efficient RST invariant feature extraction method. The Fourier descriptors based methods are applied on closed boundary shape of an object [11], which may be hard to get especially in Sindhi characters that usually are separated from each other. These characters have multiple radicals and dots (i.e. ﷽، پ ، گ, etc). The FD don't extract global information, it is due to the fact that, the inter-relationship between the contour components is considered [12]. For this reason, the FD is no more suitable for recognizing such characters. In this paper, another approach called "Sector Projection Fourier Descriptor" (SP-FD) is proposed. The SPFD is exploring the distributed regional pixel of characters using the sector projection. In the Sindhi script fonts, one of the main features is the orientation of dots. These proposed approaches have been applied on all Sindhi characters set, isolated, and characters with different scale and fonts.

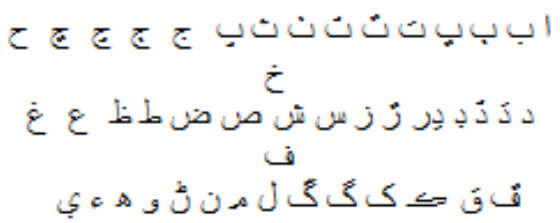


Fig. 1. Common Sindhi Characters.

TABLE. I. COMPARISON OF SINDHI CHARACTERS WITH ARABIC AND PERSIAN CHARACTERS

Sindhi Characters	Persian Characters	Arabic Characters
ب پ ٻ ٽ ٿ ٽ ٿ	ب پ ٽ ٿ	ب ٽ ٿ
ح ڃ ڄ ڃ ڇ خ	ح ڃ ڇ خ	ح ڃ خ
د ذ ڏ ڊ ڊ	د ذ	د ذ
ر ز ڑ	ر ز ڑ	ر ز
ف ق ڦ	ف ق	ف ق
ک ڪ گ ڳ گ	ک گ	ک
م ۾	م	م
ن ڻ	ن	ن
ھ ۂ ۽ ۾ ۿ	ھ	ھ

This paper is organized as follows: Section 2 describes the data preparation procedure. Section 3 explains the pre-processing technique followed by feature computation with FD and invariant characteristics. Accordingly, Section 4 presents the comparative analysis of SP-FD (Sector Projection Fourier Descriptors) with invariant feature extraction methods. Finally, Section 5 concludes the article.

II. DATA PREPERATION

The Sindhi characters contain different shapes for instance; curves, arcs, loops, junctions and lines. The isolated Sindhi characters with different scales and different fonts have been employed in this study. During the data preparation process, the original input images are prepared by manually typing the characters in Microsoft Word. The input images have been prepared in 7 different fonts and 12 different sizes. Then, the prepared are saved as ".BMP" files. All fifty six characters have been used in this study. As a result, fifty six files with seven different fonts (MB Khursheed, MB Lateefi, MB Kufi, AS Basit, MB Supreen Shabir Kumbhar Bahij Nassim and Bahij The Sans Arabic Extra Bold) have been created Subsequently, 12 different font sizes (10, 16, 20, 24, 28, 32, 36, 40, 44, 48, 56 & 72) have been used during dataset creation. The entire dataset consists of 4704 samples in total, as:

Total Samples = 56 x 7 x 12 = 4704.

These 4704 samples will yield experimental characters for all Sindhi characters consisting of different shapes, fonts and sizes. The succeeding section of the article explains the feature extraction technique used for the computation of intricate features from the input images. These features are then used for character recognition.

III. FEATURE EXTRACTION

A compact numerical representation of an object is commonly referred to as a feature vector. Contextually, the process of finding out all these unique patterns from an object of interest is known as feature extraction. In the character recognition system, the feature extraction is the most important step. For a recognition process, selection of discriminant features is an important factor. Our goal is to find a set of features that can define the shape of the underlying character as precisely and uniquely as possible [7]. There are many feature extraction techniques exists. These techniques can be categorized into the structural features, statistical feature and the global transformation methods [8, 9]. Generally, the techniques based on Fourier Descriptors are more efficient feature extraction tool used in majority of the digital image processing applications. It represents the boundary of the object by applying Fourier transform (FT) on a (character) shape that is derived from shape boundary coordinates. This is a one dimensional function and is referred to as the shape signature [10, 11]. The Fourier Descriptors are invariant under Rotation, Scale and Translation (RST) and are computationally efficient. The discrete FT is given by:

$$a_m = \frac{1}{M}\sum_{k=0}^{M-1} A(k)\exp\left(-\frac{j2\pi mk}{M}\right), \mathbf{m = 0,1,..,M-1} \quad (1)$$

Fourier coefficients a_m are chosen, the normalized magnitudes of a_n are used as features to describe the shape. A variety of FD methods depend on the closed boundary of an object's shape. However, in case of Sindhi characters; this boundary information is difficult to achieve. This is due to the fact that these characters have multiple dots and multiples radicals which are isolated from one another. For example, the Sindhi character 'پ' has five boundaries that are disconnected from each other. Another example is character '۾', which has three boundaries and all the boundaries are also disconnected from each other. Fig. 2 displays both the exemplary characters.

In order to compute the shape boundary, the proposed algorithm consists of the following steps:

Algorithm

Input: Single Sindhi character binary image

Output:

Feature Vector

(1) Find the center of mass of binary image, and set the center of mass at the origin and find the radius.

(2) Find the particular projection of shape signature by using ring projection function.

(3) To Calculate FD: Apply FT on $v(\varphi)$ and obtain the coefficients.

$$v(\varphi) = \int_0^R \delta(\rho,\varphi)d\rho,\ \varphi\ \varepsilon[0,2\pi]$$

(4) The magnitude of each coefficient is calculated and selected to obtain the feature vector.

(5) To measure dissimilarity between two dissimilar feature values, the Camberra distance method has been used.

Fig. 2. Five and Three Disconnected Closed Boundaries.

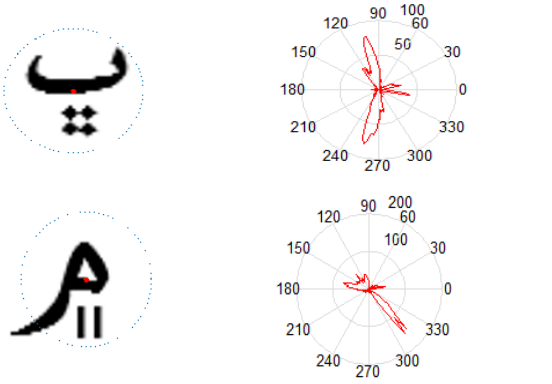


Fig. 3. Sector Projection of Two Sindhi Characters in Polar Space.

In this work, the proposed SP-FD for a particular projection, following the ring projection that can overcome the problem such as one dimensional FD method. The centre of mass (x_c,y_c) is calculated by setting the centre of mass at origin. Next, the maximum radius of the circle covering the character is obtained. The ring feature extraction function in terms of binary image character $f(x,y)$ is then defined as:

$$\delta(\rho,\varphi)=\begin{cases}1, & \text{if } g(\rho,\varphi)\in R\\ 0, & \text{Otherwise}\end{cases} \tag{2}$$

Where g(ρ,φ) represents polar form of binary image $f(x,y)$, R is the region of character's image. Then the original image will be translated at origin. Afterwards, all the pixels (counted at a particular angle φ) of image character are projected onto the series of radius denoted by $v(\varphi)$. The operation of continuous sector projection can be defined as:

$$v(\varphi)=\int_0^R \delta(\rho,\varphi)d\rho,\ \varphi\varepsilon[0,2\pi] \tag{3}$$

The discrete form of Equation (3) is:

$$v(\varphi_t)=\sum_{k=0}^{n}\delta(\rho_k,\varphi_t) \tag{4}$$

Where n represents number of samples in the radial direction, $\rho_k = k\varphi r$. Fig. 3 illustrates the sector projection for Sindhi characters 'ب' and 'م'.

After applying the Fourier transform on sector projection function, the Fourier coefficients are calculated as follows:

$$f(k)=\int_0^{2\pi} v(\varphi)\ \exp(-2\pi ik\varphi)\,d\varphi \tag{5}$$

Thus, the discretized form is:

$$f_k=\sum_{k=0}^{K-1} v\left(\frac{2k\pi}{K}\right)\exp\left(\frac{4\pi^2 ki}{K}\right) \tag{6}$$

Where K = 2π/Δφ, Δφ is angular step. To obtain the FDs, the normalized coefficients magnitudes $|f_k|$ are calculated to form a FD feature vector.

IV. Sindhi Character Recognition

Basically, the Sindhi isolated letters are built from 52 main alphabets. Each character may have different shape and differ from other characters in terms of the position of dots. The dots can be above, below, or inside the main shape of the character. In the first step, we classify the Sindhi characters according to their basic shapes. As a result, the procedure yields 56 main (isolated) forms for 52 Sindhi characters. Table II exposes 20 shapes and 20 classes. The total number of classes is 20 as given in Table III.

The characters classification is based on SP-FD and Transformation Ring Projection (TRP) feature vectors by means of statistical analysis. In order to measure the similarity and dissimilarity amongst feature vectors, the Camberra distance can be applied (as explained in Section 4). In the second step, we have enhanced the number of classes up to 56 classes, denoted by A_i, i=1,2,….,56. The relationship of two classes is Ci and Ai as shown in Table III. In this step we have used all the characters having similar shape and their respective dot position separately to find the class of characters.

TABLE. II. The 20 different Classification of Sindhi Characters

Class number	Basic shape of Sindhi character elimination of dots	Class number	Basic shape of Sindhi character elimination of dots
C_01	ا	C_11	ق
C_02	ب	C_12	ک
C_03	ح	C_13	
C_04	د	C_14	گ
C_05	ر	C_15	ل
C_06	س	C_16	م
C_07	ص	C_17	ن
C_08	ط	C_18	و
C_09	ع	C_19	
C_10	ف	C_20	ی

TABLE. III. THE RELATIONSHIP BETWEEN C AND A

Class Ci	Sub-Class Ai	Class Ci	Sub-Class Ai
C1	{A1, A56, A43}	C11	A37
C2	{A2, A3, A4, A5, A6, A7, A8, A9, A10, A46, A47, A59}	C12	A38
C3	{A11, A12, A13, A14, A15, A16, A17}	C13	A39
C4	{A18, A19, A20, A21, A22, A23}	C14	A40, A41, A42
C5	{A24, A25, A26}	C15	A43
C6	{A27, A28}	C16	A44, A45
C7	A29, A30	C17	A46, A47
C8	A31,A32	C18	A48
C9	A33,A34	C19	A49
C10	A35,A36	C20	A55

V. EXPERIMENTAL RESULTS

In Sindhi language, there are 52 main characters and depending on the position of each character in a word, it may have different shapes. Fig. 4 represents the different shapes of the same character.

Fig. 4. Different Shape of Sindhi Character.

As aforementioned, in this work only the isolated shape of each character is considered.

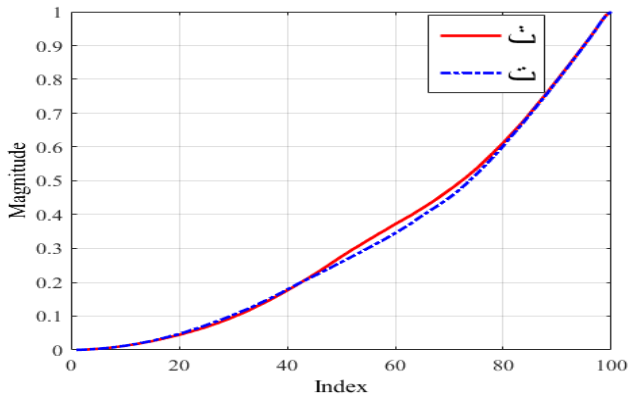
Similarity Measurement: The similarity is measured for two characters A_i and A_j, the corresponding FD feature vector are represented as $v^i = [v_1^i, v_2^i, ..., v_n^i]^T$ and $v^j = [v_1^j, v_2^j, ..., v_n^j]^T$ respectively. For the statistical analysis, the dissimilarity is measured using Camberra distance [13].

$$dissim(Ai, Aj) = \sum_{k=1}^{n} \frac{|v_k^i - v_k^j|}{v_k^i + v_k^j} \quad (7)$$

Fig.5 (a) depicts the feature extraction of two Sindhi characters ‘ٹ’and ‘ت’ using TRP method**.** These two characters possess the same shape whereas the only difference is in the position of dots. The red and blue lines display the feature values for character ‘ٹ’ and ‘ت’ respectively**.** It is evident that there exists a small difference in numeric value of feature vectors making discrimination between two characters. for a 100 dimensional vector to represent RST-invariant features 95% similarity is achieved. Therefore, it shows 5% confidence of the classifier to declare the difference between two characters. Fig. 5(b) exhibits the SP-FD features extracted for the two similar characters. Contrary to the previous methods, the similarity measurement boosts the confidence to declare that these two characters are much different. This enhanced confidence level is achieved by considering 29 feature values only. Therefore, a much better observation is presented using SP-FD features.

A. Scaling Invariance

In this case, we present the experiments performed on a scaled single Sindhi character. By scaling the character *s* times, the testing set T= {(4), (6), (8) and (10)} as shown in Fig. 6(a). The SP-FD feature vector for each entry in is shown in Fig. 6(b). It shows that the similarity among Sindhi character ‘پ’ at four different scales. The similarity achieved for character ‘پ’ is from 85-95%. These testing characters belong to the same class, and it can be said that, the method of SP-FD is verified. Fig. 6(c) illustrates the sector projection of Sindhi character ‘پ’.



(a) TRP feature extracted for Sindhi characters ‘ٹ’ and ‘ت’.

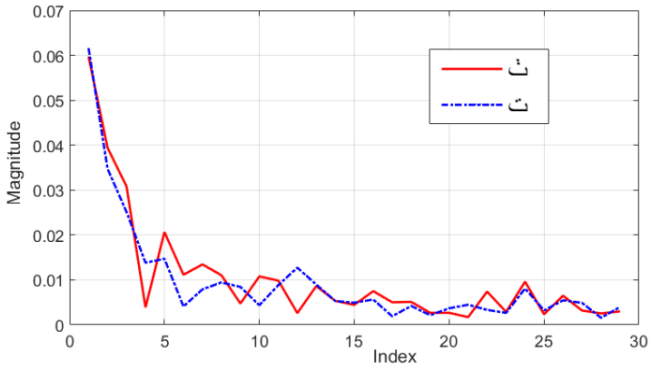
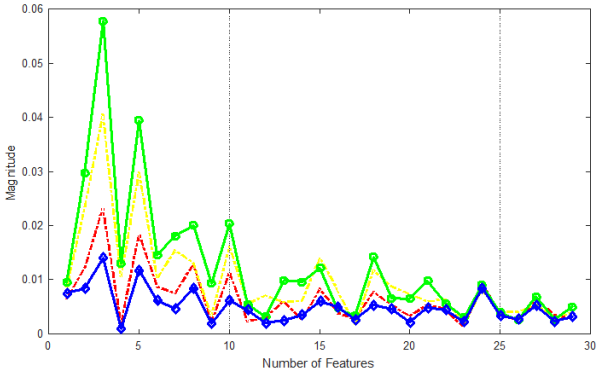


Fig. 5. (b) The Feature Extracted for Sindhi Characters ‘ٹ’ and‘ت’ using Sector Projection Fourier Descriptors.

(a) Testing Sample of Sindhi Character 'پ'.



(b) The Similarity Measured using SP-FD Algorithm from Testing Samples of Sindhi Characters 'پ'.

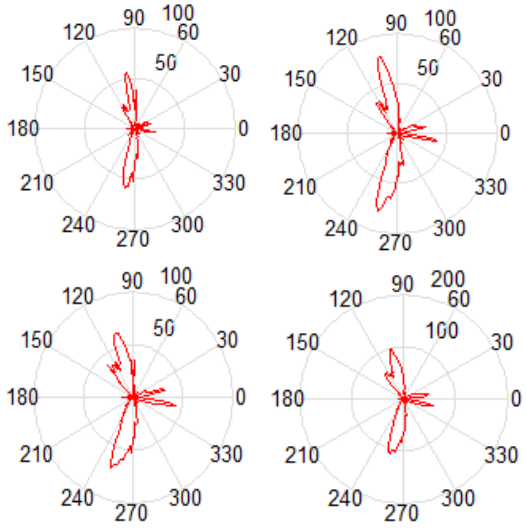


Fig. 6. (c) Projection of Four Testing Samples for پ.

VI. Conclusion

This paper presents an isolated machine printed Sindhi character recognition system. This study proposes SP-FD technique for computation of scale invariant features from Sindhi characters containing different sizes and shapes. This work also presents a comparative analysis of the proposed techniques with other techniques such as RST-invariant feature extraction methods and TRP invariant feature extraction methods. The statistical analysis clearly shows that the proposed technique outperforms other techniques proposed in the available literature. The MATLAB script only extracts the feature vector of a single image. The dissimilarity between different images can be computed by using many existing distance measurement, e.g. l-2, l-1 norm or other norms. Furthermore, the experimental results also confirm that the SP-FD feature extraction method can produce better results in Sindhi character recognition. In future, we aim to further investigate the accuracy of the proposed feature extraction method using supervised machine learning techniques such as support vector machines (SVM) and artificial neural networks (ANNs).

Acknowledgment

The authors are thankful Prof. Dr. Ahsan Ahmad Ursani, Department of Biomedical Engineering, Mehran University of Engineering & Technology, Jamshoro, Pakistan for his kind guidance.

References

[1] Impedovo, S., Ottaviano, L., and Occhinegro, S., "Optical Character Recognition: A Survey", International Journal of Pattern Recognition and Artificial Intelligence, [DOI:org/10.1142/S0218001491000041], Volume 5, No. 1, pp. 1-24, June, 1991.

[2] Trier, Ø. D., Jain, A.K., and Taxt, T., "Feature Extraction Methods for Character Recognition: A survey", Pattern Recognition, Volume 29, No. 4, pp. 641-662, [DOI: org/10.1016/0031-3203 (95)00118-2], 1996.

[3] Broumandnia, A., and Shanbehzadeh, J., "Fast Zernike Wavelet Moments for Farsi Character Recognition", Image and Vision Computing, Volume 25, No. 5, pp. 717-726, [DOI: org/10.1016/j.imavis.2006.05.014], 2007.

[4] Hakro, D.N., Ismaili, I.A., Talib, A.Z., and Mojal, G.N., "Issues and Challenges in Sindhi OCR", Sindh University Research Journal (Science Series), Volume 2, No. 46. pp. 143-152, Jamshoro, Pakistan, 2014.

[5] Abdel Raouf, A.M., "Offline Printed Arabic Character Recognition", Ph.D. Thesis, School of Computer Science, University of Nottingham, UK, 2012.

[6] Yang, M., Kpalma, K., and Ronsin, J., "Shape-Based Invariant Feature Extraction for Object Recognition", Advances in Reasoning-Based Image Processing Intelligent Systems, pp. 255-314,[DOI: 10.1007/978-3-642-24693-7_9], 2012.

[7] Shaffie, A.M., and Elkobrosy, G.A., "A Fast Recognition System for Isolated Printed Character Using Center of Gravity and Principal Axis", Applied Mathematics, Volume 4, No. 9, pp. 1313-1319, [DOI: 10.4236/am.2013.49177], 2013.

[8] Kumar, G., and Bhatia, P.K., "A Detailed Review of Feature Extraction in Image Processing Systems", IEEE 4th International Conference on Advanced Computing & Communication Technologies, pp. 5-12, [DOI: 10.1109/ACCT.2014.74], Rohtak, India,2014.

[9] Taha, S., Babiker, Y., and Abbas, M., "Optical Character Recognition of Arabic Printed Text", IEEE Student Conference on Research and Development, pp. 235-240, [DOI: 10.1109/SCOReD.2012.6518645], Pulau Pinang, Malaysia, 2012.

[10] Zhang, D., and Lu, G., "A Comparative Study of Fourier Descriptors for Shape Representation and Retrieval", Proceedings of 5th Asian Conference on Computer Vision, pp.646-651, [DOI: 10.1.1.73.5993], Springer, Melbourne, Australia,2002.

[11] El-ghazal, A., Basir, O., and Belkasim, S., "Farthest Point Distance: A New Shape Signature for Fourier Descriptors", Signal Processing: Image Communication, Volume 24, No. 7, pp. 572-586, [DOI: org/10.1016/j.image.2009.04.001], 2009.

[12] El-ghazal, A., Basir, O., and Belkasim, S. "Invariant Curvature-Based Fourier Shape Descriptors", Journal of Visual Communication and Image Representation, Volume 23, No. 4, pp. 622-633 [DOI: org/10.1016/j.jvcir.2012.01.011], 2012.

[13] Johnson, R.A., and Wichern, D.W., "Applied Multivariate Statistical Analysis", New Jersey: Prentice-Hall, 2014.

[14] Dong, L., Wang, J., Li, Y., & Tang, Y. Y. (2013, June). Sector projection fourier descriptor for chinese character recognition. IEEE International Conference on Cybernetics (CYBCO) pp. 162-167, 2013.