

F-ratio Based Weighted Feature Extraction for Similar Shape Character Recognition

T. Wakabayashi*, U. Pal**, F. Kimura* and Y. Miyake*

*Graduate School of Engineering, Mie University,

1577 Kurimamachiya-cho, TSU, Mie 514-8507, Japan

**Computer Vision and Pattern Recognition Unit, Indian Statistical Institute, Kolkata-108, India

Email: umapada@isical.ac.in

Abstract

Recognition of handwritten similar shaped character is a difficult problem and in character recognition system most of the errors occur from similar shaped characters. In this paper we proposed a novel feature extraction technique to improve the recognition results of two similar shaped characters. The technique is based on F-ratio (Fisher Ratio), a statistical measure defined by the ratio to the between-class variance and within-class variance. F-ratio modifies the feature vector of two similar shape characters by weighting the feature elements. This weighting scheme enhances the feature elements that belongs to the distinguishable portions of the similar shaped characters and reduces the feature elements of the common portion of the characters, so that similar shaped characters can be identified easily. We considered pair of handwritten similar shape characters of different scripts like Arabic/Persian, Devnagari English, Bangla, Oriya, Tamil, Kannada, Telugu etc. and we noted that f-ratio based feature weighting shows better recognition results.

1. Introduction

Recognition of handwritten characters has been a popular research area for many years because of its various application potentials. Some of its potential application areas are postal automation, bank cheque processing, automatic data entry, etc. Various approaches have been proposed by the researchers towards handwritten character recognition and many recognition systems for isolated handwritten numerals/characters in languages like English, Chinese, Japanese, Indian etc. are available in the literature [1-4]. Although high accuracy is obtained from some of the systems, it may be noted that most

of the errors are due to similar shaped handwritten characters. Recognition of these similar shaped characters is one of the difficult problems and in this paper we proposed a novel feature extraction technique to improve the recognition results of two similar shaped characters. The technique is based on F-ratio, a statistical measure that is defined by the ratio to the between-class variance and within-class variance. F-ratio is calculated from feature vectors belong to the similar shaped character classes and enhanced the feature vector for better recognition. F-ratio modifies the feature vector of two similar shape characters by enhancing the feature elements that belongs to the distinguishable portions of the similar shaped characters and reducing the feature elements of the common portion of the characters, so that these similar shaped characters can be identified easily. This is done by weighting the feature elements. To get the idea of some similar shaped characters of different scripts considered, we provided some of their similar shape printed characters in Fig.1. Although from these similar shape printed characters we can find some small differences but sometimes it is very difficult to get any difference because of writing style of different individuals.

Researchers have used many methods of feature extraction for handwritten characters. Shadow code [5], fractal code [7], profiles [8], moment [11], template [12], structural (points, primitives) [2], wavelet [9], directional feature [10] etc., have been addressed in the literature as features. From the literature survey of the existing pieces of works on characters recognition, it was evident that not much effort is given on feature enhancement to remove the confusion between similar shaped characters for their recognition. To get improved recognition results, here we introduced feature extraction based on F-ratio measure.

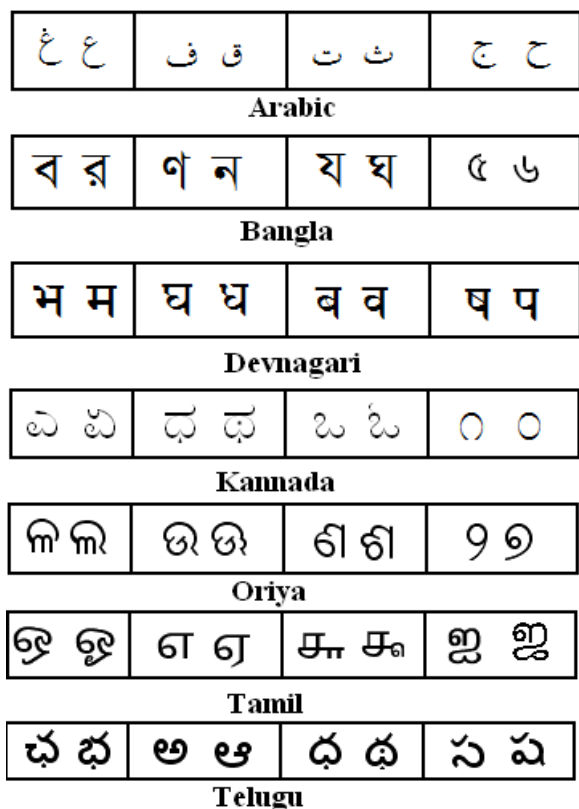


Fig.1. Examples of similar shaped characters in different scripts.

The organization of rest of the paper is as follows: In Section 2 we illustrate feature extraction techniques. Confusion character recognition is discussed in Section 3. Experimental results are described in Section 4. Finally, conclusion is presented in Section 5.

2. Feature extraction

To compare the improvement of results of the use of F-ratio we computed 400 dimensional gradient features and then we use F-ratio on this feature vector to enhance the feature by weighting, for handling similar shaped pattern. Details of 400-dimensional gradient feature and F-ratio based feature weighting are discussed as follows.

2.1 Computation of gradient feature

The gray-scale local-orientation histogram of the component is used for 400 dimensional feature extraction. To obtained 400-dimensional gradient based feature vector the following steps are executed.

Step 1: A 2×2 mean filtering is applied 5 times on the input image.

Step 2: The gray-scale image is normalized so that the mean gray scale becomes zero with maximum value 1.

Step 3: Normalized image is then segmented into 9×9 blocks. Compromising trade-off between accuracy and complexity, this block size is decided from the experiment. To get the bounding box of the grey-scale image, the image is converted into two-tone using Ostu method [6]. This will ignore unnecessary background part from the image.

Step 4: A Roberts filter is then applied on the image to obtain gradient image. The arc tangent of the gradient (direction of gradient) is quantized into 16 directions and the strength of the gradient is accumulated with each of the quantized direction. By strength of Gradient ($f(x, y)$) we mean

$$f(x, y) = \sqrt{(\Delta u)^2 + (\Delta v)^2} \text{ and}$$

by direction of gradient ($\theta(x, y)$) we mean

$$\theta(x, y) = \tan^{-1} \frac{\Delta v}{\Delta u}, \text{ where}$$

$\Delta u = g(x+1, y+1) - g(x, y)$, and

$\Delta v = g(x+1, y) - g(x, y+1)$, and

$g(x, y)$ is a gray scale at (x, y) point.

Step 5: Histograms of the values of 16 quantized directions are computed in each of 9×9 blocks [18-19].

Step 6: Directional histogram of 9×9 blocks is down sampled into 5×5 by a Gaussian filter. Thus, we get $5 \times 5 \times 16 = 400$ dimensional feature.

2.2 Feature weighting using F-ratio

There are not many differences between the feature vectors extracted from patterns of similar classes. It is difficult to classify the similar shaped pattern based on these feature vectors. To take care of such cases, we propose a feature weighting method based on F-ratio that is calculated statistically [16-17] from feature vectors belong to the similar classes. The F-ratio is computed for each element of the feature vectors and elements that have higher F-ratios include more useful information to discriminate the similar classes. Each element x_i of an original feature vector

$X = (x_1, x_2, \dots, x_n)$ is multiplied by corresponding F-ratio F_i and we obtain weighted feature vector $Z = (z_1, z_2, \dots, z_n)$ by

$$z_i = F_i x_i (i = 1, 2, \dots, n), \quad (1)$$

where n is the dimension of the feature vector. The F-ratio F_i is defined by

$$F_i = s_{bi}^2 / s_{wi}^2, \quad (2)$$

where s_{bi}^2 and s_{wi}^2 are between-class variance and within-class variance, respectively. They are defined by the following equations using class number l , class size L and *a priori* probability $P(\omega_l)$ of class ω_l .

$$\begin{aligned} s_{wi}^2 &= \sum_{l=1}^L P(\omega_l) E\{(x_i - m_{li})^2 | \omega_l\} \\ s_{bi}^2 &= \sum_{l=1}^L P(\omega_l) (m_{li} - m_{0i})^2 \\ m_{li} &= E\{x_i | \omega_l\} \\ m_{0i} &= E\{x_i\} = \sum_{l=1}^L P(\omega_l) m_{li} \end{aligned} \quad (3)$$

The function E calculates the expectation value of its argument. In our experiments, $P(\omega_l) = N_l / \sum_{l=1}^L N_l$ is used as the *a priori* probability, where N_l denotes each sample size of the similar classes.

For our experiment we consider the dimension of the original feature vector $X = (x_1, x_2, \dots, x_n)$ as 400 and this 400-dimensional feature vector is the gradient feature as computed in Section 2.1. To take care of similar shaped patterns an illustration of feature weighting done by F-ratio is shown in Fig.2. Here illustration is shown for two numerals 7 and 9, which are similar in shape. From the weighted features of 7 and 9 it can be seen that the distinguishable regions of these two similar shaped numerals gain more weights by this F-ratio (See the second and third blocks of third row of the last two images of Fig.2 and the similar portion of these similar shaped numerals loose weights. Such behavior of F-ratio weighted features helps to distinguish two similar shape patterns.

3. Character classifier

Character recognition is carried out using the following quadratic discriminant function [10]. Kimura et al. [13] compared seven statistical classifiers for handwritten zip-code numeral recognition and they obtained best results from

quadratic classifier and hence we use this classifier for our experiment.

$$\begin{aligned} g(Z) &= (N + N_0 + n - 1) \ln[1 + \frac{1}{N_0 \sigma^2} [\|Z - M\|^2 \\ &- \sum_{i=1}^k \frac{\lambda_i}{\lambda_i + \frac{N_0}{N} \sigma^2} \{\Phi_i^T (Z - M)\}^2]] + \sum_{i=1}^k \ln(\lambda_i + \frac{N_0}{N} \sigma^2) \end{aligned}$$

where Z is the feature vector; M is a mean vector of samples; Φ_i^T is the i^{th} eigen vector of the sample covariance matrix; λ_i is the i^{th} eigen value of the sample covariance matrix; k is the number of eigen values considered here, n is the feature size; σ^2 is the initial estimation of a variance; N is the number of learning samples; and N_0 is a confidence constant for σ and N_0 is considered as N . We do not use all the eigen values and their respective eigen vectors for the classification. Here, we sort the eigen values in descending order and take first 100 ($k=100$) eigen values and their respective eigen vectors for classification.

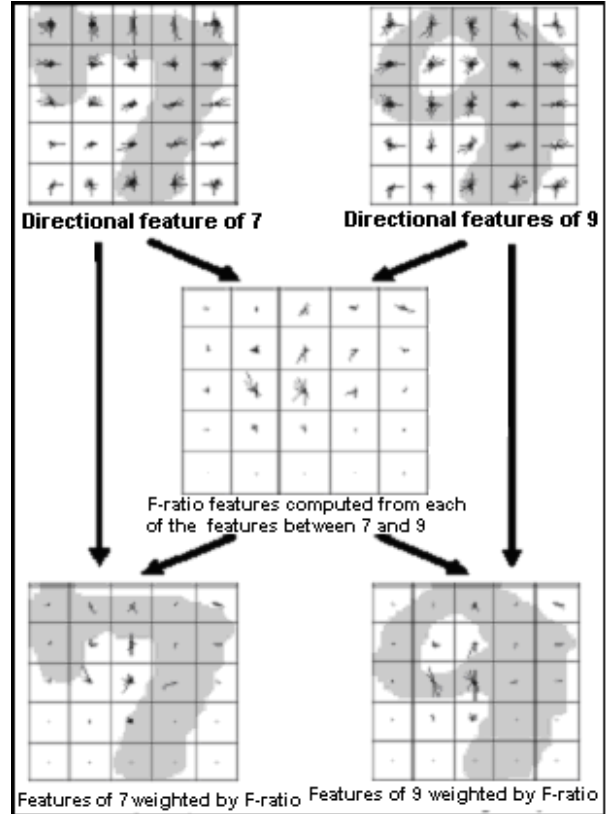


Fig.2. F-ratio weighted feature extraction for similar shaped pattern recognition. Here we consider 7 and 9 for illustration.

4. Experimental results

Before giving the detail results we will briefly discuss about the data we used for the experiment.

Data set:

For the experiment of our system we considered similar shaped characters/numerals of different scripts like English, Arabic, Devnagari, Bangla, Oriya, Tamil, Kannada, Telugu etc.

To get the similar shape numeral pair in Devnagari, Bangla, Oriya, Kannada, Tamil, Telugu and Arabic/Persian scripts we considered their 22546, 14650, 2220, 5638, 4820, 2690, 20000 samples, respectively, for the experiment [14,15]. Those numeral pair for which we get maximum errors are considered for F-ratio testing. Similarly, to get the similar shape character pair in Devnagari, Bangla, Oriya, Kannada, Tamil and Telugu scripts we considered 38859, 14879 18190, 10779, 3462 and 10873 character samples of Devnagari, Bangla, Telugu, Oriya, Kannada, and Tamil scripts, respectively, for the experiment. Those character pair for which we get maximum errors are considered for F-ratio testing.

We have used 5-fold cross validation scheme for recognition result computation. Here database of similar shape characters is divided into 5 subsets and testing is done on each subset using rest four of the subsets for learning. The recognition rates for all the test subsets are averaged to calculate recognition accuracy.

Recognition results:

Recognition results of some confusion pairs of different scripts are given in Table-1. Recognition results before and after using F-ratio are given in the table to get the idea about the improvement of F-ratio feature. From the table it may be noted that F-ratio based weighted features enhanced the recognition result of all the similar shape samples.

Error Analysis:

To get the idea about the samples where our system generates errors after using F-ratio weighted feature, we provide some erroneous samples in Table 2. Actual handwritten character samples are shown in the first row of this table and the printed samples of their recognized class are shown in the respective columns of second row. Since the actual handwritten samples and recognized characters are very similar in shape we may think that these samples are recognized correctly. Unfortunately they all mis-recognized. Actual class of each presented sample is shown in respective columns of the third row of the table.

Table 1: Recognition results on confusing pair of different scripts.

























Script	Confusing character pairs (Number of samples of each class used in the experiment is given in bracket)		Recognition Results (%)	
			Before use of F-ratio	After use of F-ratio
Devnagari	 (765)	 (784)	95.30	95.73
	 (864)	 (681)	90.55	91.06
Bangla	 (858)	 (810)	99.58	99.70
	 (166)	 (167)	88.28	90.69
English	 (3012)	 (3012)	99.07	99.27
Kannada	 (512)	 (534)	95.68	96.16
	 (484)	 (474)	99.16	99.26
Oriya	 (376)	 (358)	88.38	89.48
	 (376)	 (368)	94.33	95.14
Tamil	 (280)	 (286)	95.21	96.09
Arabic/Persian	 (8000)	 (8000)	98.90	99.83
	 (300)	 (800)	98.83	99.33

Table 2: Examples of some erroneous samples

Actual Samples (Handwritten)				
Recognized as (Printed sample)				
Actual Class (Printed sample)				

5. Conclusions

In this paper we proposed a feature enhancement technique to improve the recognition results of two similar shaped characters. The techniques is based on F-ratio. F-ratio modifies the feature vector of two similar shape characters by enhancing the feature elements that belongs to the distinguishable portions of the similar shaped characters and reducing the feature elements of the common portion of the characters, so that these similar shaped characters can be identified easily. This is done by weighting the feature elements using F-ratio. For experiment we considered pair of similar shape characters of different scripts like English, Arabic/Persian, Devnagari, Bangla, Kannada, Oriya, Tamil, Telugu etc. and we noted that F-ratio based feature weighting has ability to improve the recognition results on similar shaped characters. Here we used F-ratio only for confusion character class but F-ratio based feature can be calculated for the entire symbol set and the authors hope that proposed F-ratio based feature enhancement technique will be helpful to the researchers for their future work.

6. References

- [1] R. Plamondon and S. N. Srihari, "On-line and off-line handwritten recognition: A comprehensive survey", IEEE Trans. on PAMI, Vol.22, pp. 62-84, 2000.
- [2] U. Pal and B. B. Chaudhuri, "Indian Script Character Recognition: A Survey" Pattern Recognition, vol.37, pp.1887-1899, 2004.
- [3] A. Amin, "Off Line Arabic Character Recognition - A Survey" In Proc. 4th ICDAR, pp.596 - 599, 1997.
- [4] C. Y. Suen, S. Mori, S. H. Kim, C. H. Leung, "Analysis and Recognition of Asian Scripts: The State of the Art, In Proc. ICDAR, pp.866-878, 2003.
- [5] Harifi and A. Aghagolzadeh, "A New Pattern for Handwritten Persian/Arabic Digit Recognition", Int. Jour. of Inf. Tech., Vol. 3, pp. 249-252, 2004.
- [6] N. Otsu, "A Threshold selection method from grey level histogram", IEEE Trans on SMC, Vol.9, pp.62-66, 1979.
- [7] S. Mozaffari, K. Faez and H. Rashidy Kanan, "Recognition of Isolated Handwritten Farsi/Arabic Alphanumeric Using Fractal Codes", Image Analysis and Interpretation 6th Southwest Symposium, pp. 104-108, 2004.
- [8] J. Sadri, C. Y. Suen and T. D. Bui, "Application of Support Vector Machines for Recognition of Handwritten Arabic/Persian Digits", Proceeding of the 2nd Conference on Machine Vision and Image Processing & Applications, Vol.1, pp. 300-307, 2003.
- [9] Mowlaei, K. Faez and A. Haghghat, "Feature Extraction with Wavelet Transform for Recognition of Isolated Handwritten Farsi/Arabic Characters and Numerals" Digital Signal Processing's Vol. 2, pp. 923-926, 2002.
- [10] F Kimura, K. Takashina, S. Tsuruoka and Y. Miyake, "Modified quadratic discriminant function and the application to Chinese character recognition", IEEE Trans. on PAMI, Vol. 9, pp 149-153, 1987.
- [11] M. Dehghan and K. Faez, "Farsi Handwritten Character Recognition With Moment Invariants", Proceedings of 13th International Conference on Digital Signal Processing, Vol. 2, pp. 507-510, 1997.
- [12] M. Ziaratban, K. Faez and F. Faradji "Language Based Feature Extraction Using Template-Matching in Farsi/Arabic Handwritten Numeral Recognition", In Proc. 9th ICDAR, pp. 297-301, 2007.
- [13] F. Kimura, S. Nishikawa, T. Wakabayashi, Y. Miyake and T. Tsutsumida, "Evaluation and synthesis of feature vectors for handwritten numeral recognition", IEICI Trans. Inf. and System, Vol.E79-D, pp.436-442, 1996.
- [14] H. Khosravi, E. Kabir, "Introducing a very large dataset of handwritten Farsi digits and a study on the variety of handwriting styles", Pattern Recognition Letters Vol.28, Issue 10, pp. 1133-1141, 2007.
- [15] U. Pal, T. Wakabayashi, N. Sharma and F. Kimura, "Handwritten Numeral Recognition of Six Popular Indian Scripts", Proc. 9th ICDAR, pp. 749-753, 2007.
- [16] A. Blum and P. Langley, "Selection of relevant features and examples in machine learning", Artificial Intelligence, Vol.97, pp. 245-271, 1997.
- [17] X. He, D. Cai and P. Niyogi, "Laplacian Score for Feature Selection", Advances in Neural Information Processing Systems, vol. 18, pp. 49-156, 2005.
- [18] U. Pal, T. Wakabayashi and F. Kimura, "Handwritten Bangla Compound Character Recognition using Gradient Feature", In Proc. 10th International Conf. on Info. Tech., pp. 208-213, 2007.
- [19] M. Shi, Y. Fujisawa, T. Wakabayashi, and F. Kimura, "Handwritten numeral recognition using gradient and curvature of gray scale images", Pattern Recognition, Vol.35, pp.2051-2059, 2000.