Statistical Features Extraction for Indian Language Documents

[1]Manoj Kumar Shukla, [2]Haider Banka
[1]Amity School of Engineering, Amity University, Noida, India
[2]Dept. of CSE, Indian School of Mines, Dhanbad, Jharkhand, India

Abstract
India is a multi-lingual, multi script country. Therefore developing a successful multi-lingual OCR, system for feature extraction of different scripts is a very important step. In this paper we discussed a Statistical Features based algorithm for feature extraction. A family of procedures for measuring relevant shape information in a pattern in order to make the task of classifying pattern easy is called Feature Extraction. It analyses a segment of text and selects features which are unique to the text and can be used to identify it. The soul of pattern recognition system design is selection of features that are stable and representative.

Keywords
Feature Extraction, Indian Language Documents, Multi-lingual OCR, Pattern Recognition

I. Introduction

Feature extraction assumes a significant part in the fruitful distinguishment of machine-printed and transcribed characters [1-2]. Characteristic extraction could be characterized as the methodology of concentrating dissimilar data from the frameworks of digitized characters. In OCR requisitions, it is significant to concentrate those characteristics that will empower the framework to separate between all the character classes that exist. Numerous distinctive sorts of characteristics have been recognized in the expositive expression that may be utilized for character and numeral distinguishment.

Two primary classifications of features are Global (measurable) and Structural (topological) [2]. Worldwide features are those that are concentrated from each purpose of a character lattice. At first, some worldwide systems were intended to recognize machine-printed characters [3]. Worldwide features might be distinguished all the more effectively and are not as delicate to neighborhood clamor or contortions as are topological features. Nonetheless, in a few cases minor measure of commotion may have an impact on the real arrangement of the character framework, subsequently relocating features. This may have genuine repercussions for the distinguishment of characters influenced by these mutilations [3, 2]. Worldwide features themselves may be further separated into various classifications. The predominant and most straightforward feature is the state of every last one of focuses in a character lattice. In a parallel picture there are just dark or white pixels, the state consequently alludes to if a pixel is dark or white. One methodology that has been for the most part utilized for extraction of worldwide features is dependent upon the factual circulation of focuses [1]. Six systems that have been utilized in the literary works, in light of the appropriation of focuses, are quickly sketched out in next sub-area.

Trier et al. [4] summarized and analyzed a percentage of the well-known feature extraction routines for disconnected from the net character distinguishment. Determination of a feature extraction strategy is likely the single most vital element in realizing high distinguishment execution in character distinguishment frameworks. They talked over feature extraction strategies as far as invariance lands, reconstructability and needed contortions and variability of the characters.

In addition the factual and structural features, arrangement extension coefficients are likewise utilized as features of a character. The expositive expression on these three classes is talked about beneath.

II. Statistical Features

Statistical Features have been additionally used to concentrate characteristics from sectioned debased characters. We have utilized the accompanying factual feature.

A. Zoning

The extracted character picture (row or re-scaled), is portioned into windows of equivalent size. Thickness values (Number_ of_foreground_pixels/ Total_number_of_pixels) are acquired for every window. All thickness qualities are utilized to structure the data characteristic vector for a specific character design. As characterized by Trier et al. [4], zoning might be characterized as the methodology in which a n × m matrix is superimposed on the character picture and for each of the n × m zones; the normal ash levels if there should be an occurrence of light black level character pictures are utilized as gimmicks. If there should be an occurrence of double pictures the rate of dark pixels in each one zone is registered [4]. As zoning is definitely not invariant to scaling, we have scaled the characters before discovering peculiarities utilizing zoning.

B. Moments

Moments are unadulterated factual measure of pixel dissemination around middle of gravity of characters and permit catching worldwide character shapes data. They portray numerical amounts at some separation from a reference point or hub. They are intended to catch both worldwide and geometric data about the picture. Minute based invariants investigate data over a whole picture instead of giving data exactly at single limit point, they can catch a portion of the worldwide properties lost from the immaculate limit based representations like the general picture introduction. In 1962, Hu [5] utilized the hypothesis of mathematical invariants and inferred his seven acclaimed invariants to revolution of 2-D articles. Since that time, minute invariants have turned into an established apparatus for peculiarity based article distinguishment. The first Hu's invariants used the second and third-request minutes just. The development of the invariants from higher-request minutes is not direct. Fourier-Mellin change, Zernike polynomials and mathematical invariants have been utilized within a number of uses as given in subtle elements by Teh and Chin [6] to accomplish invariant distinguishment of two-dimensional picture designs.

Geometric moments or general moments are easy to execute. Hu's invariants have the attractive properties of being invariant under picture interpretation, scaling, and revolution. Notwithstanding, it is discovered that to figure the higher request of Hu's minute invariants is truly complex, and to reproduce the picture from Hu's invariants is likewise extremely troublesome. A pioneer take a shot

at this field was carried out freely by Reiss [7] in 1991 and by Flusser and Suk [8] in 1994. They redressed a few oversights in Hu's hypothesis, presented relative minute invariants (Ami's), and demonstrated their pertinence in basic distinguishment undertakings. Hu [5] and Flusser and Suk [8] have demonstrated that a set of eleven invariant capacities have been generally utilized. Zernike initially proposed the Zernike polynomials in 1934. Their minute detailing seems, by all accounts, to be one of the most prevalent, beating theplan B (as far as clamor flexibility, data repetition and remaking ability). Complex Zernike moments have been widely utilized as the invariant worldwide gimmicks for picture distinguishment. Zernike moments have been investigated and actualized by Teh and Chin [6] and Khotanzad and Hong [9]. Singh [10] has utilized skimming point computations for Zernike moments.

C. Zernike Moments

Zernike presented a set of complex polynomials {vnm(x, y)} which structure a complete orthogonal set over an unit plate of x2 + y2 ≤ 1. The manifestation of the polynomial is:

$$V_{nm}(x,y) = R_{nm}(x,y)e^{jm\theta} \quad (1)$$

where $j=\sqrt{-1}$, $\theta = \tan^{-1}\left(\frac{y}{x}\right)$

and

$$R_{nm}(x,y) = \sum_{s=0}^{n-|m|/2} \frac{(-1)^s (x^2+y^2)^{(n-2s)/2}(n-s)!}{s!((n+|m|-2s)/2)!((n-|m|-2s)/2)!} \quad (2)$$

where $n \geq 0$, $(n-|m|)$ = even, and $|m| \leq n$.

For a Digital image the Zernike moments of order n and repetition m are expressed as:

$$A_{nm} = \frac{n+1}{\pi} \sum_{x=0}^{M-1} \sum_{y=0}^{N-1} f(x,y) V_{nm}^{*}(x,y), \quad (3)$$

Where where $x^2 + y^2 \leq 1$ and * denotes the complex conjugate operator.

The characterized gimmicks of Zernike minutes are just invariant to revolution. To attain scale what's more interpretation invariance, the picture needs to be standardized by utilizing normal Zernike moments.

The interpretation invariance is accomplished by deciphering the original picture.

$f(x, y)$ to $f(x+\bar{x}, y+\bar{y})$, where $\bar{x} = \frac{m_{10}}{m_{00}}$ and $\bar{y} = \frac{m_{01}}{m_{00}}$

The original picture's inside is moved to the centroid before the Zernike moments figuring.

Scale invariance is accomplished by expanding or lessening or scaling the picture about its

centroid and after that performing the scaling operation to change over the picture to a standard size. A standout amongst the most paramount properties of the Zernike moments is its energy to remake the picture from the figured moments. Assume we know all Zernike moments Anm of f(x, y) up to request N. Because of the orthogonal property of Zernike moments, we can recreate the picture in light of this set of Zernike moments by:

$$\hat{f}(x,y) = \sum_{n=0}^{n_{max}} \sum_{m=-n}^{n} A_{nm} V_{nm}(x,y) \quad (4)$$

where n_{max} is the maximum order of the Zernike moments considered.

III. Orthogonal Fourier Mellin Moments

OFMMs, presented by Sheng and Shen [11], hold more nearby data about character distinguishment, which are focused around a situated of spiral polynomials. They could be utilized to speak to little characters in the same path as huge characters. As depicted by Kan and Srinath [12], the amount of OFMMs needed to speak to a picture is much lower than that of ZMs so OFMMs might be more vigorous than ZMs if the characters have expansive variability also portray little picture all the more faultlessly when expansive size character specimens are

taken for preparing and moderately littler size characters are utilized for testing.As Shen and Sheng presented OFMMs, the round Fourier or Radial Mellin Moments (FMMs) of a picture capacity $f(r, \theta)$ are characterized in the polar direction framework (r,θ) as

$$M_{s,m} = \int_0^{2\pi} \int_0^{\infty} r^s f(r,\theta) e^{-jm\theta} r\,dr\,d\theta \quad (5)$$

where $f(r, \theta)$ is the picture and m = 0, + 1, + 2, ... is the roundabout consonant request. By

definition, the Mellin change request s is intricate esteemed. With whole number s ≥ 0, OFM moments presently might be characterized as:

$$\phi_m = \frac{1}{2\pi a_n} \int_0^{2\pi} \int_0^{1} f(r,\theta) e^{-jm\theta} r\,dr\,d\theta \quad (6)$$

where an is a standardization steady and Qn(r) is a polynomial in r of degree n. The set of Qn(r) is orthogonal over the extent 0 ≤ r ≤ 1:

$$\int_0^1 Q_n(r) Q_k(r) r\,dr = a_n \delta_{nk} \quad (7)$$

where δnk is the Kronecker image and r = 1 is the greatest size of the protests that can be experienced in a specific application. Subsequently the premise capacities $Q_n(r)\ e^{-jm\theta}$ of the OFMM are orthogonal over the inner part of the unit loop. OFMM might be considered summed up ZM. They have a solitary orthogonal set of the outspread polynomials $Q_n(r)$ for all round consonant request q, while ZM have one orthogonal set of spiral polynomials $R_p^{|q|}(r)$ for every diverse round request q.

ZM concentrate on the global featurs and get less nearby data than OFMM. Turn, interpretation and scale invariance could be gotten in the same route as with Zernike moments proposed by Kan and Srinath [9].

IV. Acknowledgments

In this present paper we have discuss Statistical Features Extraction algorithm for different type of Indian language document like (Devnagari and Bangla) . In this type of work there are very wide research scopes. The research work should be done to enhance the documents containing these kind of extraction of feature in Indian language script document and subsequently recognition them. The

algorithm used in the paper is very simple, easy to understand and reliable for the line wise segmentation of the scripts. In the algorithm, the segmentation rate of the scripts is very fast and accurate.

References

[1] S. Impedovo, L. Ottaviano, S. Occhinegro,"Optical character recognition- A survey", International Journal Pattern Recognition and Artificial Intelligence, Vol. 5(1-2), pp. 1-24, 1991.

[2] C. Y. Suen,"Character recognition by computer and applications", In Handbook of Pattern Recognition and Image Processing, New York: Academic, pp. 569-586, 1986.

[3] C. Y. Suen, M. Berthod, S. Mori,"Automatic recognition of hand printed characters- the state of the art", Proceedings of the IEEE, Vol. 68(4), pp. 469-487, 1980.

[4] O. D. Trier, A. K. Jain, T. Taxt,"Feature extraction methods for character recognition - A survey", Pattern Recognition, Vol. 29(4), pp. 641-662, 1996.

[5] M. K. Hu,"Visual pattern recognition by moment invariants", IRE Transactions on Information Theory, Vol. 8(2), pp. 179-187, 1962.

[6] C. H. Teh, R. T. Chin,"On image analysis by the method of moments", IEEE Transactions on PAMI, Vol. 10(4), pp. 496-513, 1988.

[7] T. H. Reiss,"The revised fundamental theorem of moment invariants", IEEE Transaction on PAMI, Vol. 13(8), pp. 830-834, 1991.

[8] J. Flusser, T. Suk,"Affine moment invariants: a new tool for character recognition", Pattern Recognition Letters, Vol. 15(4), pp. 433-436, 1994.

[9] A. Khotanzad, Y. H. Hong,"Invariant image recognition by Zernike moments", IEEE Transactions on PAMI, Vol. 12(5), pp. 489-497, 1990.

[10] C. Singh,"Improved quality of reconstructed images using floating point arithmetic for moment calculation", Pattern Recognition, Vol. 39(11), pp. 2047-2064, 2006.C.

[11] Y. Sheng, L. Shen,"Orthogonal Fourier-Mellin moments for invariant pattern recognition", J. Optical Society of America, Vol. 11(6), 1748-1757, 1994.

[12] C. Kan, M. D. Srinath,"Invariant Character Recognition with Zernike and Orthogonal Fourier-Mellin Moments", Pattern Recognition, Vol. 35, pp. 143-154, 2002.