Statistical Feature Extraction Methods for Isolated Handwritten Gurumukhi Script

Dharam Veer Sharma, Geeta Saini, Mohit Joshi
(Assistant Professor, Department of Computer Science, Punjabi University, Patiala.)
(M.Tech (Computer Science) Department of Computer Science, Punjabi University Patiala.)
(M.Tech (Computer Science) Department of Computer Science, Punjabi University Patiala.)

ABSTRACT

In this Paper we have used moment based Statistical methods for Feature extraction like Zernike, Pseudo-Zernike methods. Handwritten Gurmukhi isolated characters are used for feature extraction. Our database consists of 75-115 samples of each of 40 characters of Gurmukhi script collected from different writers. These samples are pre-processed, normalized and scaled to 48*48 sizes. Feature extraction at various moment orders has been done. Reconstruction of the sample images are done to check the accuracy of the computed features.

Keywords: Statistical features, Zernike, Pseudo-Zernike, order, moments, isolated handwritten Gurmukhi character.

1. INTRODUCTION

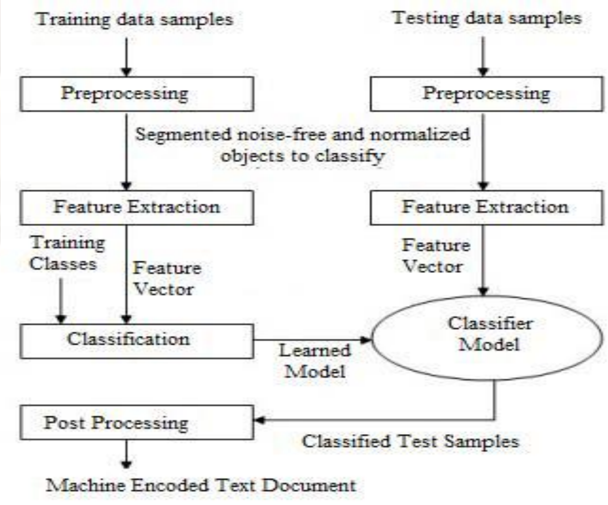
Feature Extraction is a special form of dimensionality reduction. When the input data to an algorithm is too large to be processed and it is suspected to be notoriously redundant (much data, but not much information) then the input data will be transformed into a reduced representation set of features (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.

Different feature extraction methods are designed for different representations of the characters, such as solid binary characters, character contours, skeletons, or gray level sub images of each individual characters. The feature extraction methods are discussed in terms of invariance properties, reconstructability, noise sensitivity, computational complexity[1]. Extraction of good features is the important phase to correctly recognize an unknown character. A good feature set contains discriminating information, which can distinguish one object from other objects. It must also be as robust as possible in order to prevent generating different feature codes for the objects in the same class. The selected set of features should be a small set whose values efficiently discriminate among patterns of different classes, but are similar for patterns within the same class. Feature extraction is the process of selection of the type and the set of features. Features can be classified into two categories:

1**. *Local features,*** which are usually *geometric/structural.* Structural features are involved of structural elements like loop, line, crossing point, branches, joints, curve, end points and stroke etc. Various structural Feature extraction methods are in presented in [2].

2. ***Global features,*** which are usually *topological/statistical.* Statistical features are computed by some statistical operations on image pattern and these include features like zoning, projection, profiling, histogram and distance, moments etc. Methods and strategies for statistical feature extraction in handwritten script identification was presented in [1][3].

Structural and statistical features appear complementary to each other and many other features can be derived from the basics of these features. The basic mechanism of offline character recognition consists of following phases: *Image Pre-processing, Feature Extraction, Classification and Post Processing*. (Fig.1) [4]



Gurmukhi script is used primarily for the Punjabi language, which is the world's 14th most widely spoken language. Punjabi is an Indo-Aryan language spoken by about 105 million people mainly in West Punjab in Pakistan and in East Punjab in India. In India, Punjabi is written with the Gurmukhi alphabet, while in Pakistan it is written with a version of the Urdu alphabet known as Shahmukhi. The written standard for Punjabi in both India and Pakistan is known as Majhi, which is named after the Majha region of Punjab. Punjabi is one of India's 22 official languages and it is the first official language in East Punjab. The Gurmukhi alphabet developed from the Landa alphabet and was standardized during the 16th century by Guru Angad Dev Ji, the second Sikh guru. The name Gurmukhi means "from the mouth of the Guru" and comes from the Old Punjabi word *guramukhi*.Gurmukhi script is cursive. There is rich literature in this language in the form of scripture, books, poetry. It consists of 35 basic characters; there are 10 vowels and modifiers, 6 additional modified consonants, forming 41 consonants

ੳ	ਅ	ੲ	ਸ	ਹ	ਕ	ਖ	ਗ	ਘ	ਙ	
ਚ	ਛ	ਜ	ਝ	ਞ	ਟ	ਠ	ਡ	ਢ	ਣ	
ਤ	ਥ	ਦ	ਧ	ਨ	ਪ	ਫ	ਬ	ਭ	ਮ	
ਯ	ਰ	ਲ	ਵ	ੜ	ਸ਼	ਜ਼	ਖ਼	ਫ਼	ਗ਼	ਲ਼
ੌ	ੇ	ੋ	ੈ	ੱ	ਁ	ਃ	ਿ	ੀ	ਾ	
ੁ	ੂ	਼	੍							

(figure 1.2)

Figure 1.2 Character set of Gurmukhi Script

Features of Gurumukhi Script:

- Direction of writing: left to right in horizontal lines.
- Type of writing system: syllabic alphabet (consist of symbols for consonants and vowels).
- Consonants have an inherent vowel. Diacritics, which can appear above, below, before or after the consonant they belong to, are used to change the inherent vowel.
- When certain consonants occur together, special conjunct symbols are used which combine the essential parts of each letter.

1.3 Shape descriptors

Shape is one of the fundamental visual features in the content-based image retrieval (CBIR) paradigm. Shape descriptor is some set of numbers that are produced to describe a given shape feature. A descriptor attempts to quantify shape in ways that agree with human intuition. Good retrieval accuracy requires a shape descriptor to be able to effectively find perceptually similar shapes from a database. Usually, the descriptors are in the form of a vector. Shape descriptors should meet the following requirements:

- The descriptors should be as complete as possible to represent the content of the information items.
- The descriptors should be represented and stored compactly. The size of descriptor vector must not be too large.
- The computation of distance between descriptors should be simple; otherwise the execution time would be too long.

Shape feature extraction and representation plays an important role in the following categories of applications:

- *Shape retrieval*: searching for all shapes in a typically large database of shapes that are similar to a query shape.
  Usually all shapes within a given distance from the query are determined or the first few shapes that have the smallest distance.

- *Shape recognition and classification*: determining whether a given shape matches a model sufficiently or which of representative class is the most similar.
- *Shape alignment and registration*: transforming or translating one shape so that it best matches another shape, in whole or in part.
- *Shape approximation and simplification*: constructing a shape of fewer elements (points, segments, triangles, etc.) that is still similar to the original.

Many shape representation have been proposed for various purposes. These methods can generally be grouped into:

- contour-based/boundary based
- region based

Contour-based shape descriptors make use of only the boundary information, ignoring the shape interior content. Therefore, these descriptors cannot represent shapes for which the complete boundary information is not available. Contour-based methods, such as chain code, shape signature, polygonal approximation, autoregressive models and Fourier Descriptor ; exploit shape boundary information which is crucial to human perception in judging shape similarity. On the other hand, region-based descriptors exploit both boundary and internal pixels, and therefore are applicable to generic shapes. Region based methods, such as geometric moments, Zernike moments, grid representation and area, exploit only shape interior information, therefore can be applied to more general shapes. Some geometric features such as average scale, skew, kurtosis, etc. reflected in the region based methods are also important perceptual features.

2. MOMENTS

An *image moment* is a certain particular weighted average (moment) of the image pixels' intensities, or a function of such moments, usually chosen to have some attractive property or interpretation. Image moments are useful to describe objects after segmentation. Image normalization should be used prior to moment extraction for applications requiring invariance.Simple properties of the image which are found *via* image moments include area (or total intensity), its centroid, and image orientation.

Moment based feature descriptors have evolved into a powerful tool for image analysis applications. Moments of the gray value-function *f*(*x*, *y*) of an object can be defined as:

$$m_{p,q} = \int\int x^p y^q f(x,y) dx dy$$

The integration is calculated over the area of the object. Pixel based features instead of the gray value could be used to calculate the moments of the object.

2.1 Order of Moments

Moments are generally classified by the order of the moments. The order of a moment depends on the indices *p* and *q* of the moment *m(p, q) and* vice versa. The sum *p* + *q* of the indices is the order of the moment *m(p,q).*
For Example:

- zero order moment ((*p*, *q*) = (0, 0))

The zero order moment describes the area *A* of the object as:

$$m_{0,0} = \int\int dx dy b(x,y)$$

- first order moments ((*p*, *q*) = (1, 0) or (0, 1))

The first order moments contain information about the center of gravity of the object

$$m_{1,0} = \int\int dx dy x f(x,y)$$
$$m_{0,1} = \int\int dx dy y f(x,y)$$

- second order moments ((*p*, *q*) = (2, 0) or (0, 2) or (1, 1)))

$$m_{2,0} = \int\int dx dy x^2 f(x,y)$$
$$m_{0,2} = \int\int dx dy y^2 f(x,y)$$
$$m_{1,1} = \int\int dx dy x y f(x,y)$$

The moments are features of the object, which allow a geometrical reconstruction of the object. To get more precise description of complex objects, moments of higher order and more complex moments like Zernike or Pseudo-Zernike moments are to be used. Higher order moments should be used to get a minimal error reconstructing object by image moments.

3. ZERNIKE MOMENTS

Moments of orthogonal polynomial basis were proposed by Teague [5]. They have proven less sensitive to noise, are natively invariant to linear transformations and can be effectively used for image reconstruction. Computational complexity, however, becomes a major issue. Geometric moments present a low computational cost, but are highly sensitive to noise. Furthermore, image reconstruction is extremely difficult. Moments of discrete orthogonal basis have been proposed .They are fast to implement, present adequate noise tolerance and very accurate image reconstruction. Image normalization should be used prior to moment extraction for applications requiring invariance.

A set of orthogonal functions with simple rotation properties which forms a complete orthogonal set over the interior of the unit circle was introduced by Zernike. The form of these polynomials is:

$$V_{nm}(x,y) = V_{nm}(\rho sin\theta, \rho cos\theta) = R_{nm}(\rho) exp(jm\theta) \quad (3.1)$$

where n is either a positive integer or zero and m takes positive and negative integers subject to constraints n - |m| = even, |m| ≤ n, ρ is the length of the vector from the origin to the pixel at (x, y), and θ is the angle between vector ρ and the x axis in the counterclockwise direction. The Radial polynomial $R_{n,m}(\rho)$is defined as:

$$R_{nm}(\rho) = \sum_{s=0}^{(n-|m|)/2} (-1)^s \frac{(n-s)!}{s!\left(\frac{n+|m|}{2}-s\right)!\left(\frac{n-|m|}{2}-s\right)!} \rho^{n-2s}, \quad (3.2)$$

with $R_{n\ -m}(\rho) = R_{n\ m}(\rho)$.

Zernike moments present native rotational invariance and are far more robust to noise. Scale and translation invariance can be implemented using moment normalization.

The complex Zernike moments of order n with repetition m for an image function *f(x,y)* can be defined as:

$$A_{nm} = \frac{n+1}{\pi}\int\int_{x^2+y^2\leq 1} f(x,y)\, V^*_{nm}(\rho,\theta)\, dx\, dy, \tag{3.3}$$

or, in polar coordinates

$$A_{nm} = \frac{n+1}{\pi}\int_0^{2\pi}\int_0^{1} f(\rho,\theta)\, R_{nm}(\rho)\, exp(-jm\theta)\, \rho\, d\rho\, d\theta. \tag{3.4}$$

Where real valued radial polynomial $R_{n,m}(\rho)$ is defined in (3.2)

Due to the conditions n - |m| = even and |m| ≤ n for the Zernike polynomials in (3.1) the set of Zernike polynomials contains ½ (n+1)(n+2) linearly independent polynomials if the given maximum degree is n.

Zernike moments descriptor does not need to know boundary information, making it suitable for more complex shape representation. Zernike moments descriptors can be constructed to arbitrary order, this overcomes the drawback of geometric moments in which higher order moments are difficult to construct.

4. PSEUDO-ZERNIKE MOMENTS

If we eliminate the condition n - |m| = even from the Zernike polynomials defined in (3.1) $\{V_{n\ m}\}$ becomes the set of pseudo-Zernike polynomials. The set of pseudo-Zernike polynomials was derived by Bhatia and Wolf [7] and has properties analogous to those of Zernike polynomials.

For the pseudo-Zernike polynomials, the real -valued radial polynomial $R_{n\ m}(\rho)$ is defined as:-

$$R_{nm}(\rho) = \sum_{s=0}^{n-|m|} (-1)^s \frac{(2n+1-s)!}{s!\,(n-|m|-s)!\,(n+|m|+1-s)!}\rho^{n-s}, \tag{4.1}$$

Where n=0, 1, 2,…, ∞ and m takes on positive and negative integers subject to |m| ≤ n only. Unlike the set of Zernike polynomials this set of pseudo-Zernike polynomials contains $(n+1)^2$ linearly independent polynomials instead of ½ (n+1) (n+2) if the given maximum order is n.

The definition of the pseudo-Zernike moments is the same as that of the Zernike moments in (3.3) and (3.4) except that the radial polynomials $\{R_{n\ m}(\rho)\}$ in (4.1) are used. Since the set of pseudo-Zernike orthogonal polynomials is analogous to that of Zernike polynomials.

5. RESULTS AND ANALYSIS

Zernike and Pseudo-Zernike methods are used for feature extraction in our OCR system. About 4068 samples of different isolated characters of Gurmukhi script has been used for feature extraction. Features are extracted at different orders. Zernike moments are computed at order 25,30,35,40,48. Best reconstruction of sample images is analyzed at order 48. Pseudo-Zernike moments are computed at order 10,15,20,24. The feature extracted at order 24 was quite good. Work is going on to improve the accuracy and speed and extensive experiments are being performed. However, the method for classification K-NN will be used for comparing the results with previous OCRs.

6. REFERENCES

[1] O. D. Trier, A. K. Jain and T. Text, “Feature Extraction Methods for Character Recognition- A Survey”, *Pattern Recognition*, Vol. 29, No. 4, pp. 641-662, 1996.

[2] G.S.Lehal and Chandan Singh,” Feature extraction and classification for OCR of Gurmukhi script”, Vivek, Vol. 12, No. 2, pp. 2-12 (1999).

[3] Dharamveer Sharma, Puneet Jhajj, “Recognition of Isolated Handwritten Characters in Gurmukhi Script”, *International Journal of Computer Applications* (0975-8887), Vol. 4, No. 8, 2010.

[4] K.S. Siddharth, M.Jangid, Renu Dhir, R.Rani , ”Handwritten Gurumukhi character recognition using Statistical and Background Directional Distribution Features”, Vol. 3 No. 6 June 2011.

[5] M. R. Teague, "Image analysis via the general theory of moments", J. Opt. Soc. Amer., vol. 70, pp. 920-930, Aug. 1980.

[6] Simon Xinmeng Liao, “Image Analysis by Moments”, Thesis[Online]Available:http://zernike.uwinnipeg.ca/~s_liao/pdf/thesis.pdf.

[7] A.B Bhatia and E. Wolf, Proc. Camb. Phil. Soc., Vol.50 pp.40-48, 1954.

[8] Wikipediawebsite:Imagemoment[Online]Available:http://en.wikipedia.org/wiki/Image_moment.

[9] L. Kotoulas and I. Andreadis, “Image analysis using Moments”.Volume: 1, Issue: 1, Publisher: Citeseer Pages: 360–364.

[10] Wikipedia website-Feature Extraction [Online] Available:http://en.wikipedia.org/wiki/Feature_extraction.

[11] P. B. Saad , Feature Extraction of Trademark Images Using Geometric Invariant Moment and Zernike Moment – A Comparison, Vol.31 No.3 (DECEMBER 2004).