Indian Journal of Science and Technology, *Vol 8(13), 56453, July 2015*
ISSN (Print) : 0974-6846
ISSN (Online) : 0974-5645

Structural Features Extraction for Devnagari and Bangla Language Documents

Manoj Kumar Shukla[1]*, Haider Banka[1] and K. P. Yadav[2]

[1]Department of Computer Science and Engineering, Indian School of Mines, Dhanbad - 826004, Jharkhand, India; mkshukla001@gmail.com, haider.banka@gmail.com
[2]Department of Computer Science & Engineering, Sunder Deep Group of Institutions, Ghaziabad - 201001, U. P., India; drkpyadav732@gmail.com

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 Structural 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. The real objective of feature extraction is to concentrate a situated of features, which expands the distinguishment rate with the slightest measure of components and to produce comparable feature set for mixture of cases of the same image. Feature extraction systems investigate the data archive picture and select a set of features that extraordinarily distinguishes and characterizes the character. Feature Extraction and Classification procedures are essential steps in character distinguishment methodology to attain to high distinguishment execution. Feature extraction is characterized as the issue of "separating from the crude information the data which is most important for characterization purposes, in the feeling of minimizing the inside class design variability while upgrading the between-class design variability". In this paper we have developed structural features for Devnagari and Bangla script document.

Keywords: Devnagari and Bangla Language Document, Features, OCR

1. 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[2,3]. 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

*Author for correspondence

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; reconstruct ability 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.

2. Structural Features

Structural features may be characterized as far as character strokes, character openings, or other character properties, for example, concavities and convexities, end focuses and intersections, extrema, convergence with straight lines and so on. Structural features could be grouped into two classes. Lee and Chen[5] have spoken to every Chinese character by a set of short line sections, where every line fragment is spoken to by it's begin and end focus arranges. The accompanying three features are then concentrated to speak to a line fragment: the middle focus arrange, the incline and the relationships between the line section and its neighboring line sections.

Amin[6] has utilized seven sorts of structural features, for example number of sub words, number of tops of every sub word, number of circles of every top, number and position of complimentary characters, the tallness and width of every crest for distinguishment of printed Arabic content.

Lee and Gomes[7] have utilized the structural features for transcribed numeral distinguishment, for example number of focal, left and right pits, area of every focal pit, the intersection groupings, the amount of convergences with the key and auxiliary tomahawks and the pixel dissemination.

Rocha and Pavlidis[8] have proposed a technique for the distinguishment of multifont printed characters utilizing the accompanying structural features: arched bends and strokes, independent focuses and their relationships. The independent focus is one of the accompanying places: a limb focus, a closure focus, a raised vertex and a sharp corner.

In a prototypal paper Kahan et al.[9] have improved a structural list of capabilities for distinguishment of printed content of any font and measure. The list of capabilities incorporates the accompanying data for a character: number of openings, area of gaps, concavities in the skeletal structure, intersections of strokes, endpoints in the vertical heading and bouncing box of the character.

1. Local features which are normally geometric. (e.g., sunken/arched parts, number of endpoints, limbs, joints and so on.)
2. Global features, which are typically topological (connectivity, projection profile, number of holes, *etc.*)

Structural features have the accompanying points of interest.

1. It is instinctive, implying that the planner of a structure technique has full control over the parameters and the fine points of interest of the procedure. Interestingly, the results and the execution of a factual strategy, depend intensely on the parameters, features set utilized, and the preparing set.
2. It can adjust for substantial varieties in the information. The structural methodology has the ability to manage vigorously contorted information.
3. Structural techniques could be intended to exploit the entire shape meaning of an info character. The factual methodologies look just at predefined peculiarity vectors, which give just halfway data about the shape.

Structural features should to be picked remembering that the shape varieties should to influence list of capabilities insignificantly. It was not a simple errand to choose which structural features should to be decided to concentrate the

structural features from corrupted characters of Devnagari script and Bangla script because of extensive shape varieties in characters of the same class. Gimmick codes of the structural features set have after basic qualities.

- These structural features are less touchy to character size and text style.
- The gimmick codes introduce a high distinguishableness for diverse characters. In other words, the gimmick codes speaking to distinctive characters have a low likelihood to match.
- These features are really tolerant to commotion.
- Features are less delicate to character varieties, because of textual style contrasts or filtering.

We have utilized the accompanying structural features of Devnagari and Bangla characters for developing feature vector.

1. **Presence of Sidebar (St1):** This feature is available if a vertical sidebar, of roughly the same tallness as of the character, is available at the rightmost side of the sub-image. As talked about in Chapter 4, we have effectively utilized this feature for segmentation reason. Further, it is noted that if full sidebar exists in Devnagari and Bangla characters, it is dependably at the rightmost side of the character. There are 21 characters in middle zone having full sidebar at their right end. These Devnagari characters are: अ, ख, घ, च, ज, झ, ञ, ण, त, थ, ध, न, प, ब, भ, म, य, ल, व, ष, स and there are 16 characters in middle zone having full sidebar at their right end. These Bangla characters are:খ, গ, ঘ, ণ, থ, ধ, ন, প, ব, ম, য, ল, শ, ষ, স, য়.
2. **Presence of Half Sidebar (St2):** This feature is available if a sidebar, of give or take half the stature of the full character is available at the rightmost side of the sub-image. As examined in Chapter 4, we have officially utilized this feature for segmentation reason. There are 5 characters in Devnagari script having this feature: ङ, ट, ड, ढ, द.
3. **Presence of Headline (St3):** The vicinity of feature in the sub-image is an alternate imperative feature for arrangement. Actually when the sub-images are exceptionally corrupted, this feature is held. For instance, in Bangla character ঙ has no headline feature while ঘ has headline feature. In Devnagari character set all character having this feature. There are 26 characters in Bangla script having this feature: ক, ঘ, চ, ছ, জ, ঝ, ট, ঠ, ড, ঢ, ত, দ, ন, ফ, ব, ভ, ম, য, র, ল, ষ, স, হ, ড়, ঢ়, য়.
4. **Number of Junctions with Headline (St4):** It might be noted that each one character in middle zone of Devnagari and Bangla character set has either one or more than one junctions with the feature. For example in Devnagari character त has one junction with feature while क has two junctions with feature. This feature is genuine if a sub-image of Devnagari and Bangla have one junction with headline else it is false. There are 14 sub-images in Devnagari characters set having this feature genuine: ङ, च, ज, ञ, ट, ठ, ड, ढ, त, द, न, प, य, र. And there are 10 sub-images in Bangla characters set having this feature true:খ, গ, চ, ড, ণ, ন, ব, য, ল, স.
5. **Number of Junctions with the Baseline (St5):** This feature is valid for a sub-image if number of junctions with the pattern is one else it is false. This feature in Devnagari script is valid for subsymbol ढ and false for sub-image न since it has two junctions with the baseline. There are 22 subsymbols in Devnagari characters set having this feature genuine: क, ग, घ, च, ज, ञ, ट, ठ, ड, ढ, त, थ, द, ध, प, य, र, ल, व, ह. And there are 12 sub-images in Bangla characters set having this feature true: খ, গ, ঘ, চ, ড, ণ, থ, ন, ব, য, ল, স.
6. **Aspect ratio (St6):** The Aspect ratio is a main element of characters with comparable quadrant thickness as it speaks to the introduction of the character. The viewpoint ratio is the ratio between the length and the width of the character. We have separated the entire sub-images introduce in center zone into three classifications relying on the aspect ratio of the sub-images. We consider St6 = 0 if the perspective ratio is short of what 0.92.There are two more extensive characters घ and ध in Devnagari script having St6 = 0. Likewise if aspect ratio is more noteworthy than 3.0 then St6 = 2. Additionally, the estimation of aspect ratio has been utilized as a part of instance of upper zone and lower zone sub-images.
7. **Left, Right, Top and Bottom Profile Direction Codes (St7, St8, St9 and St10**): A variety of chain encoding is utilized on left, right, top and base profiles. For discovering the left profile direction codes, the left profile of a sub-image is examined

start to finish and nearby direction of the profile at every pixel is noted. Beginning from current pixel, the pixel separation of the following pixel in east, south or west direction is noted. The total number of development in three headings is spoken to by the rate events as for the aggregate number of pixel development and put away as a 3 part vector with the three parts speaking to the separation secured in east, south and west direction, individually.

8. **Directional Distance Distribution (St11):** Directional Distance Distribution (DDD) is a distance based feature proposed by Oh and Suen[10]. For each pixel in the information double show, two sets of 8 bytes which are called W(white) set and B(black) set are apportioned. For a white pixel, the set W is utilized to encode the distances to the closest dark pixels in 8 direction (0°, 45°, 90°, 135°, 180°, 225°, 270°, 315°). The set B is basically loaded with quality zero. Additionally, for a dark pixel, the set B is utilized to encode the distances to the closest white pixels in 8 directions. The set W is filled with zeros.
9. **Transition Features (St12):** In this structural feature, area and number of transitions from foundation to closer view pixels in the vertical and even Horizontal are noted. The transition feature utilized here is like that proposed by Gader et al.[11]. To compute transition data, picture is checked from left-to-right, right-to-left, start to finish and bottom to-top. To guarantee a uniform feature vector estimate, the moves in every course are processed as a small amount of the separation crossed over the picture. For instance, if the transitions were being figured through and through, a transitions discovered near the top would be relegated a high esteem contrasted with a move processed further down. A greatest worth (M) was characterized to be the most extreme number of moves that may be recorded in every heading.

On the other hand, if there were short of what M moves recorded (n for case), then the staying M - n moves would be doled out estimations of 0 (to support in the formation of uniform vectors).It will deliver four matrices, two matrices having measurements NC × 5 and other two matrices having measurements NR × 5 (where NC speaks to the Number of Columns/width of the character matrices and NR speaks to the quantity of Rows/tallness of the Character), and we the second phase of transition feature computation comprises of resampling the transition locations onto altered size matrices. For that, we have isolated each grid on a level plane into T equivalent amounts of. We have taken the normal moves vertically in each one section. At last, if NC = 50, NR = 50, M = 5 and T = 5, we got a 4 × 5 × 5 feature vector.

3. Conclusion

In this present paper we have discuss Structural 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.

4. References

1. Impedovo S, Ottaviano L, Occhinegro S. Optical character recognition - a survey. International Journal Pattern Recognition and Artificial Intelligence. 1991; 5(1-2):1–24.
2. Suen CY. Character recognition by computer and applications. Handbook of Pattern Recognition and Image Processing. New York: Academic; 1986. p. 569–86.
3. Suen CY, Berthod M, Mori S. Automatic recognition of hand printed characters - the state of the art. Proceedings of the IEEE. 1980; 68(4):469–87.
4. Trier OD, Jain AK, Taxt T. Feature extraction methods for character recognition: - a survey. Pattern Recognition. 1996; 29(4):641–62.
5. Lee HJ, Chen B. Recognition of handwritten Chinese characters via short line segments. Pattern Recognition. 1992; 25(5):543–52.
6. Amin A. Recognition of printed Arabic text based on global features and decision tree learning techniques. Pattern Recognition. 2000; 33:1309–23.
7. Lee LL, Gomes NR. Disconnected handwritten numeral image recognition. Proceedings of 4th ICDAR; 1997. p. 467–70.
8. Rocha J, Pavlidis T. A shape analysis model with applications to a character recognition system. IEEE Transactions on PAMI. 1994; 16(4):393–404.
9. Schenkel M, Jabri M. Low resolution, degraded document recognition using neural networks and hidden