Degraded Script Identification for Indian Language- A Survey

Manoj Kumar Shukla
Research Scholar
Department of CSE
ISM-Dhanbad

Haider Banka, Ph.D.
Assistant Professor,
Department of CSE
ISM-Dhanbad

ABSTRACT
The working module of any Optical character Recognition system almost depends upon printing and paper of the input document image. A number of OCR techniques are available and claim correctly identified accuracy in printed document image in Indian and foreign script. A few report have been found on the recognition of the degraded Indian language document. The degradation in any scanned printed document can be of many types. In this paper, we focus a survey of degraded script identification for Indian Language document.

Keywords
OCR, Degraded Script.

1. INTRODUCTION
Despite its age, the state of symbolization in the field could achieve the purpose of pragmatic use as of late just. The procedure of OCR begins with perusing of an examined picture of an arrangement of characters, verifies their importance, and at long last deciphers the picture to a workstation composed content report. Additionally, organizations and citizens can utilize this system to rapidly interpret paper records to PC composed reports.

A mess of examination has been carried out on OCR in most recent 55 years. A few books [17-19] and numerous studies [3, 4, 20-38] have been distributed on the character distinguishment. The majority of the distributed take on OCR has been on Latin characters, with chip away at Japanese and Chinese characters developing amidst 1960s. Functional surveys and studies in the field of OCR incorporate the chronicled audit of OCR techniques and business frameworks by Mori [3], Mantas [20], Govindan and Shivaprasad [21] and Suen et al. [22]. The review by Impedovo et al. [23] keeps tabs on business OCR frameworks, while the work by Tian et al. [24] studies the territory of machine-printed OCR. Jain et al. [25] summarized and analyzed a percentage of the well-known routines utilized within different phases of an example distinguishment framework. They have tried to recognize research points and provisions which are at the front line in this field. Buddy and Chaudhuri [4] in their overview report summarized diverse frameworks for Indian dialect scripts distinguishment They have depicted some business frameworks like Tamil and Kannada OCR. They reported the extent of anticipated work to be enlarged in numerous bearings, for example OCR for low quality records, for multi font OCR and bi-script/multi-script OCR improvement and so on. A book reference of the fields of OCR and report investigation is given in [26]. [28] overviewed on-line penmanship distinguishment and depicted a contortion tolerant shape matching technique. [22] proposed an overview on systems utilized for on-line distinguishment of hand-printed characters while Connell et al. [30, 31] portrayed on-line character distinguishment for Devanagari characters and alphanumeric characters.

Bortolozzi et al. [32] have distributed an exceptionally suitable study on later developments in penmanship distinguishment. Lee et al. [33] depicted disconnected from the net distinguishment of completely unconstrained written by hand numerals utilizing multiplayer bunch neural system. The character locales are dead set by utilizing projection profiles and topographic characteristics separated from the light black scale pictures. At that point, a nonlinear character division way in every character district is considered by utilizing multi-stage chart look calculation.

Khaly and Ahmed [34], Amin [35] and Lorigo&Govindraju [36] have generated a complete study and reference index of examination on the Arabic optical content distinguishment. Hildebrandt and Liu [37] have reported the developments in manually written Chinese character distinguishment and Liu et al. [38] have talked over different procedures utilized for on-line Chinese character distinguishment.

2. SEGMENTATION
In archive investigation, division is the equivalent word for line, word or character division. Consistent with Casey and Lecolinet [39], division focuses ought not singularly be made on the foundation of neighborhood data. Rather, past and anticipated division choices may as well likewise be made on the groundwork of context oriented data. At the end of the day, if division is made between two primitives in an expression and the coming about letters don't fit with any letter stages in a saying dictionary, such division may be regarded wrong and might impact past and anticipated divisions. Casey and Lecolinet [39] summarises the most recent forty years of character division inquire about as being reliant on nearby topological data and also worldwide context oriented data. A couple of reviews have been distributed in the written works on the particular theme of character division [39-42]. Dunn and Wang [40] have partitioned the division strategies inspected in their paper into straight division and division distinguishment classes. Lu and Shridhar [42] have examined the division of hand-printed statements, written by hand numerals and cursive manually written expressions. Casey and Lecolinet [39] have partitioned their overview into four classes: analyzation systems, distinguishment based division, blended procedures (over segmentation) and all encompassing methods as talked over underneath

(a) Dissection techniques: These allude to procedures for division that are dependent upon the thought of dividing character pictures into sub-segments using general characteristics. No arrangement, context oriented learning or character shape separation steps are available in these division methods.

(b) Recognition-based segmentation: These systems don't utilize particular analyzation methods. Rather the picture is divided into covering areas and a classifier is utilized to perform division by checking if a specific area comprises of a character. This sort of procedure is alluded to as distinguishment based on the grounds that the character division is a by-result of distinguishment.

(c) Hybrid strategies (over-segmentation): These techniques are the consolidation of the first two that are specified previously. Dismemberment is utilized to over-section the expression or associated character part into a sufficient number of parts as to include all division verges present. In the following step, arrangement is utilized to verify the best divisions from a set of conceivable division speculations.

(d) Holistic strategies: At last, these methodologies expect to recognise expressions as whole units instead of endeavoring to concentrate distinctive characters.Starting frameworks for portioning machine-printed characters were dependent upon two straightforward characteristics: white space and pitch [39]. White space alludes to holes between printed characters that could be located by vertically checking the picture. A segment in the picture that didn't hold any dark pixels (frontal area pixels) could be recognized a white space. The pitch identifies with machine print provisions that yield characters of altered width. It is the amount of characters that involve a given region in the even course. Pitch data could accordingly be utilized to verify the area of just as separated division focuses in a line of printed data. This information could be viable in part smudged or consolidated character limits. Hoffman and Mccullough [43] additionally used pitch estimation of printed characters on top of an assessment capacity dependent upon a tally of dark white and white-dark moves to gauge division locales.

An alternate ubiquitous technique utilized as a part of recognizing division focuses is that of projection examination [39]. Specifically, vertical projection or vertical histograms are utilized to count the number of dark pixels in every vertical segment in the printed picture. It has served such purposes as the recognition of white spaces or ranges of low pixel thickness in printed matter. It can additionally be utilized to gauge the vicinity of vertical strokes in machine print. Baird et al. [44] used the vertical projection to verify division zones in covering or touching characters. They computed the degree of the histogram bend's second subsidiary to its stature. Minima in the histogram might show up as crests accompanying proportion computation and subsequently would demonstrate prospective part focuses. [46] proposed different strategies, in view of the vertical histogram. Wang and Jean [47] proposed a technique to section touching characters utilizing neural systems. These were based the examination exhibited via Baird et al. [44]. All these division methods have a place with the Dissection classification without a distinguishment module to affirm or fortify character part.

Utilizing distinguishment based division procedure, Casey and Nagy [48] advanced a recursive part calculation that uses a window to sweep a printed picture from left to right, testing all division conceivable outcomes. In any case, a window is set over the whole info picture. The window is then steadily contracted from the right. At every phase of narrowing, the model classifier endeavors to recognize the substance of the window. This proceeds until the recognizer makes a match with a character or the window comes to be too modest. Provided that a character is effectively recognized, the window is moved surrendered over to the point of truncation and the window is reset to at the end of the day start narrowing from the amazing right.

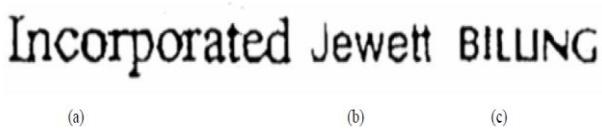
A standout amongst the most normally discovered issues in debased machine-printed records is presence of touching characters. The point when two neighboring characters touch one another, they are called combined characters or touching characters. Division of touching characters is an imperative and challenging undertaking. In the accompanying sub-areas, we have surveyed different procedures utilized within writing for sectioning touching characters/numerals.

2.1 Segmentation of Touching Characters

Division of touching characters (at times alluded to as composite characters or united characters) is a troublesome issue in character division. There are two key issues included in this issue, the first is to verify which portions hold different characters, i.e., recognizing the applicant of division, and the second is to find soften areas up hopeful of division for portioning touching characters [48].

The strategies for fragmenting the united characters might be partitioned into two classifications [41], characteristic based and distinguishment based. In the characteristic based strategy, vertical projection is changed to a capacity which furnishes more regulate data for uncovering the break focuses. The second system includes division and distinguishment process conveyed concurrently for dividing and recognizing touching characters.

The most regularly utilized characteristic for discovering different character portions (hopeful of division) is the width and the viewpoint proportion of the character. Lu [41] recommended that character width could be powerfully assessed throughout the division handle. Each one competitor fragment is then inspected by contrasting its width and the assessed character width or by measuring the viewpoint degree of the section. It is well grasped that generally characters have widths more modest than their tallness and a solitary character fragment may as well have a width of less than twice of the assessed character width. The combo of these two estimations works well more often than not yet comes up short in uncommon cases. Figure 2.1.(a) shows samples of touching character portions, "In", "rp", "ed", that are more extensive than single characters in the picture, along these lines they could be recognized utilizing either the section width or the aspect ratio.



(a) (b) (c)

Figure 2.1: Touching characters in Roman script (Dunn *et al.* [40])

Be that as it may, in Figure 2.1(b) touching character section "tt" is narrower than "J" and "w", and touching character fragment "LI" is no more extensive than "N" and "G" in the Figure 2.1(c). Besides, the angle proportions of "tt" and "LI" are as typical as a solitary character fragment. Lu [41] further proposed producing single and different character profile displays for separating multi character portions from single character sections.

In the wake of uncovering the applicant of division, the following issue is to uncover the break area which will section touching characters. It is a part procedure to confirm where to

part the various characters district. Casey and Lecolinet [39] have talked about a couple of systems for uncovering these break areas in their paper. Kahan et al. [52] proposed a destination capacity for finding breaking focuses inside the united characters. The capacity is the degree of the second contrast of the vertical projection capacity V(x) to V(x+1), to be specific,

$$f(x-1,x,x+1)=\frac{V(x-1)-2\times V(x)+V(x+1)}{V(x)} \qquad (2.1)$$

The maxima of this destination capacity were utilized as the conceivable breakpoints. Liang et al. [46] proposed a segregation capacity for uncovering the break areas for sectioning touching characters dependent upon both pixel and profile projections. Lu [41] prescribed a top to valley capacity to enhance the above strategies. The entirety of distinctions between least quality and the crests on every side is computed. The proportion of the aggregate to the base quality itself (in addition to 1, probably to escape division by zero) is the discriminator used to select division limits. This degree shows an inclination for low valley with high tops on both sides.

Despite the fact that the greater part of the scientists have utilized these systems for character division, and asserted sensible division precision, however these strategies have a few impediments. Thus, there is a need to advance new routines for degradation.

Various calculations have been proposed in the past for dividing touching characters in Roman script [46-56]. Liang et al. [46] proposed a segregation capacity for portioning touching characters dependent upon both pixel and profile projections. They proposed a dynamic recursive division calculation with a correctness of 99.40-99.85% for portioning touching characters. Wang and Jean [47] proposed a half breed technique for machine-printed character partition. They used a basic form following calculation to at first differentiate the characters and connected a neural classifier to recognise the distinctive characters. In the situation where two characters were not accurately divided by the basic dismemberment plot, further division plans were conjured. Lee et al. [49] have sectioned touching characters utilizing projection profiles and topographic characteristics separated from the light black scale pictures. At that point a nonlinear character division way in every character division area is considered by utilizing multi-stage diagram look calculation. At last, distinguishment based division strategy is utilized to affirm the division ways. Tsujimoto and Asada [50] developed a choice tree for determining vagueness in portioning touching characters. Creators utilized distinguishment come about to recognize multi character parts. Provided that a segment had a similitude measure more terrific than from the earlier edge, it was a solitary character part, any other way it was a multi-character segment. Casey and Nagy [48] proposed a recursive division calculation for dividing touching characters. Hong [51] has used visual between word stipulation accessible in a content picture to part word pictures into pieces for portioning corrupted English dialect characters. In 1987, Kahan et al. [52] have proposed a destination capacity, the portioning goal capacity, for finding breaking focuses inside blended characters. Bose and Kuo [53] utilized a vigorous structural investigation procedure dependent upon line contiguousness chart for dividing touching characters. Noticeable stroke and their headings were noted and orchestrated in climbing request. Two covering strokes with slants of inverse sign is recognized division focus. Schenkel and Jabri [54] have utilized joined division and distinguishment strategy (interior division) for producing provisional character. A space removal neural system is prepared to produce character probabilities. The yield of the neural system is then post-handled by a Hidden Markov Model that successfully seeks through the distinguishment character applicants for optical character recognizable proof and limits. Zhao et al. [55] proposed a two-stage approach to fragment unconstrained written by hand Chinese characters. In their methodology, first a character string is coarsely portioned dependent upon the vertical projection and foundation skeleton, and the squares of associated characters are distinguished; then in the fine division arrange the joined characters are divided with an exactness of 81.6%. Lu [41] inferred a crest to valley capacity to enhance the above strategies. Casey and Lecolinet [39] have talked about a couple of systems for finding break areas in touching characters. Taking into account proliferation and contracting courses of action, a calculation for dividing touching characters has been talked about by Nakamura et al. [56].

2.2 Segmentation of Touching Numerals

Numerous calculations [57-63] have likewise been proposed to fragment touching transcribed numerals. Elnagar and Alhajj [57] proposed a diminishing based calculation for portioning single touching manually written digits, in light of foundation and shape offers in conjunction with a set of heuristics to verify the potential division focuses with a correctness of 96%. Donggang and Hong [58, 59] have advanced a technique to differentiate single touching manually written numeral strings utilizing structural characteristics. In their methodology, in view of the structural focuses in the written by hand numerals strings, touching area of touching segment is dead set, and after that dependent upon the geometrical data of an unique structural focus, a hopeful touching focus is chosen. At long last, they utilized morphological examination and incomplete distinguishment comes about for the reason. Chi et al. [60] proposed a form bend based calculation to section single- and twofold touching manually written digit strings. Lu et al. [61] proposed a foundation diminishing approach for the division of joined manually written digit strings. Buddy et al. [62] have utilized water repository thought behind portioning unconstrained manually written joined numerals. They have recognized the area, size and touching position (beat, center or base) of the supply and after that investigated the store border, touching position and topological characteristics of touching example, confirming the best division focus with a correctness of 94.8%. Chen and Wang [63] utilized diminishing based technique to section single- or numerous touching manually written numerals. They performed diminishing of both forefront and foundation districts on the picture of joined numerals strings. The closure and fork focuses acquired by diminishing are utilized for cutting focuses extraction.

2.3 Segmentation of Touching Characters in Indian Scripts

Not many calculations have been researched on fragmenting touching characters in Indian scripts [64-69]. Bansal and Sinha [64] have divided the conjuncts (one sort of touching examples) in Devanagari script utilizing the structural lands of the script. Conjuncts are regularly discovered in Devanagari script pages. Conjuncts are essentially a consolidation of a half character accompanied by a full character. Both the half

and full characters dependably touch one another. So these sort of touching characters are because of structural lands of the script, yet not because of different explanations which prepares touching characters. In the strategy proposed by the creators, first the left half character of the conjunct is divided utilizing the structural lands of the Devanagari script. They suggested that the half character is discovered in one-third to half parcel of the full width of the conjunct. Beginning from one-third section of the aggregate width of the conjuncts, at whatever point one gets more number of pixels in a segment then the past segment, that section denote the half character limit. Right away, by utilizing the idea of caved in flat projection, the left limit of the second constituent character of the conjunct (which is a full character) is discovered. As asserted by the creators, the victory rate of these calculations for sectioning the conjuncts is 84%. The restriction of this work is that it will work just to fragment the conjuncts and not the composite characters.

Garain and Chaudhuri [65] have utilized a strategy dependent upon fluffy multifactorial dissection to portion touching characters in Devanagari and Bangla scripts. They proposed a prescient calculation for successfully selecting conceivable cut segments for fragmenting touching characters. The creators have asserted 98.92% correctness for effectively fragmenting touching characters. The limit of their methodology of fragmenting touching characters is the recommendation that at touching position the width of touching blob is constrained to not many sections. This strategy works fine if the width of touching blob is modest yet it doesn't work that well when the blob width at touching position is equivalent to or more stupendous than the stroke width. Prior, Garain and Chaudhuri [66] have proposed three stage controls to section touching characters in printed Bangla script. To start with, to recognize touching characters, perspective degree examination has been carried out on top of distinguishment score. The segment for which a standardized size invariant comparability measure is less than limit quality, it is suspected to be touching characters, gave that its bouncing box angle proportion is more than a predefined edge. Second, to choose about the cut position level of center ness of touching position and thickness of the single dark run is assessed and isolating these two weights for the potential cut positions is assessed. The cut position competitors are discovered at nearby maxima in the histogram of weights. Off and on again, it might generate over division. Third, to dispose of the over division each one fragmented characters is passed through distinguishment handle. Provided that the character is recognized then acknowledged, any other way this portion is united with the following fragment and again the distinguishment process is connected, on the recently built section and this process proceeds until distinguishment stage acknowledges it as a legitimate character. The creators have asserted an exactness of 97% of portioning touching characters, utilizing this calculation.

Lehal and Singh [67, 68] have given a calculation to fragment touching characters in upper zone of Gurmukhi script. Chaudhuri et al. [69] have utilized the rule of water flood, from a supply, to portion touching characters in Oriya script. The rule is that provided that we pour water on the highest point of a character, the positions where water aggregates are acknowledged as supplies. Figure 2.2 shows the repository shaped in a solitary character and in a couple of touching characters. A store whose stature is modest and which lies in the upper part of the center zone of a line is recognized as a hopeful supply for touching character division. The cusp (lowermost focus) of the supply of the competitor store is acknowledged as the detachment purpose of touching characters. On account of the round state of the majority of the Oriya characters, it was watched, that a supply is usually structured when two characters touch one another.

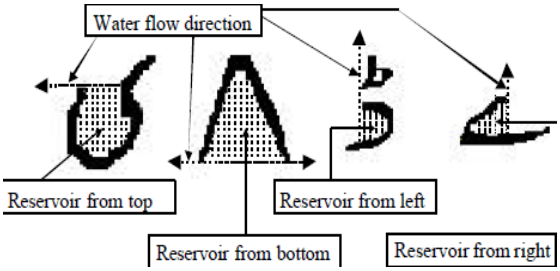


Figure 2.2: Water reservoirs (shown by dotted lines)

Not many works have additionally been accounted for dividing the evenly covering lines in Indian scripts. Bansal [70] has talked about a two-pass calculation based upon normal line stature to tackle the issue of on a level plane covering lines in Devanagari script. Harikumaret al. [71] have utilized the thought of normal line stature to section the on a level plane covering lines in Malayalam script. For sectioning unconstrained transcribed content lines of Bangla script, Pal and Datta [72] have partitioned the content into vertical strips and after that taken the level projections.

Buddy et al. [73] have utilized different characteristics of Indian scripts like presence of feature, number and position of tops in even projections, water repository and so on to differentiate different lines from multi script archive. Buddy and Chaudhuri [74] have utilized structural and measurable characteristics for differentiating machine-printed content lines from manually written message lines for both Bangla and Devanagari scripts. Dholakia et al. [75] have utilized inclines of associated parts to uncover three zones in the printed Gujarati script. Buddy and Chaudhuri [10, 76] have additionally examined the thought of zoning and line division. As per Lehal [77], the issue of sectioning different strips in Gurmukhi archives might be fathomed utilizing normal center strip stature.

3. FEATURE EXTRACTION

Feature extraction assumes a significant part in the fruitful distinguishment of machine-printed and transcribed characters [23, 78]. 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) [78]. Worldwide features are those that are concentrated from each purpose of a character lattice. At first, some worldwide systems were intended to recognise machine-printed characters [22]. 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 [22, 78]. 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 [23]. 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. [79] 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.

3.1 Statistical or Global Features

Measurable features are factual measures of appropriation of focuses on the bitmap, the shape bend, the profiles, or the HV-projections. Broadly utilized routines are minutes, zoning, n-tuples, projections, trademark loci and intersections and separations.

(a) Moments: Minute invariants are features dependent upon measurable minutes of characters. Various routines in this classification use the minutes of pixels in a picture as features. They are customarily utilized apparatuses for character distinguishment [80-83]. Traditional minute invariants were presented by Hu [84]. Hu's seven minutes are well known to be invariant to position, size and introduction of the character. They are immaculate measurable measures of the pixel appropriation around the core of gravity of the character and permit catching the worldwide character shape data. Higher request minutes are troublesome to apply. Numerous creators have proposed choices to Hu's minutes (Radial and calculated minutes [85] and Zernike minutes [86]). An alternate sort of minutes, specifically, focal minutes are ascertained by considering the separation of focuses from the centroid (focus of gravity) of the character [22, 23]. In this case, focal minutes are wanted to crude minutes as they transform higher distinguishment rates and are invariant to the interpretation of the picture [23]

(b) Zoning: This technique isolates the character lattice into minor windows or zones. The densities of focuses in every window are computed and utilized as features to the picked classifier. It was presented in 1972 by Hussain et al. [87] and has been utilized by Bosker [12] for the business OCR framework Calera and additionally by Messelodi and Modena[88].

***(c) n*-tuples:** This technique basically utilizes the event of dark or white pixels in a character picture as features. The n-tuple strategy has been investigated by Tarling and Rohwer [89]. Features removed by the n-tuple plan measure arbitrary lands of pixels.

(d) Projections: Projections are determined from histograms of Horizontal and vertical projections of dark pixels in some specific range of the character. Projections were presented in 1956 in a fittings OCR framework by Glauberman [90] and have been utilized by Heutte et al.

(e) Characteristic loci: In this strategy, horizontal and vertical vectors are produced for every white foundation pixel in a picture. Characteristics are produced by numbering the amount of times a line fragment is crossed in the horizontal and vertical bearing.

(f) Crossings and distances: Lastly, researchers have obtained features by analyzing the number of times the character image is crossed by vectors in certain directions or angles, *i.e.*, 0^o, 45^o, 90^o*etc.*[22, 23].

3.2 Structural Features

Structural features portray an example as far as its topology and geometry by giving its worldwide and neighborhood lands. A portion of the principle structural features incorporate features like number and crossing points between the character and straight lines, gaps and sunken curves, number and position of finish focuses and intersections [80]. These features are ordinarily hand created by different creators for the sort of example to be grouped. Chinese characters hold rich structural data, which remains unaltered over font and estimate variety. Since the essential components of a Chinese character are strokes, the sorts and amounts of strokes and relationships around the strokes are crucial structural features of a Chinese character. Lee and Chen [92] have spoken to every Chinese character by a set of short line sections, where every line fragment is spoken to by its 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 [93] 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 [94] 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 [95] 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. [52] 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.

3.3 Series Expansion Coefficients

Need of diminishing the extent of characteristic vectors while rendering the characteristics resistant to revolution and interpretation accelerates talk of the third worldwide characteristic class: Transformations and Series extensions. The systems for discovering characteristics by Transformations and arrangement extensions have turned out to be invariant to scaling, turn and interpretation, while decreasing dimensionality of the characteristic vector.

Numerous scientists have utilized conversions and arrangement extensions as characteristics for the assignment of character recognition [21-23, 78]. The most widely recognized arrangement development coefficient built factual characteristics depend with respect to Fourier convert [97]. In Fourier change the characters are edge and their outskirts are concentrated. The outskirt of every character might be spoken to by its Fourier convert. Fourier coefficients with critical qualities are called Fourier descriptors and might be utilized as characteristics for character recognition. The most clear provision of the Fourier change is to straightforwardly perform the two-dimensional form of the convert on the picture and take the coefficients. The coefficients or all the more frequently blends of coefficients, of these terms are utilized as characteristics.

4. CLASSIFICATIONS

Characterization is the second part of OCR motor as illustrated in Chapter 1. As recently clarified, arrangement is the part of the recognition framework that endeavors to catch the class that a specific character fits in with. After a classifier may do this, it must be demonstrated an expansive number of preparing examples, in a sort of studying stage [23]. It has been noted that the way to high execution is through the capability to select and use the different characteristics of characters. No straightforward plan is prone to accomplish high recognition rate, henceforth more complex frameworks have been created. We have talked over in this area different sorts of arrangement routines that have been investigated. Rundown of different characterization systems incorporates pattern matching, syntactic routines, factual techniques, fake neural systems, bit strategies and half breed classifiers [25, 32, 98].

4.1 Template Matching

Template matching is one of the least complex and soonest methodologies to patter recognition. In pattern matching, a model or a model of the example to be recognized is accessible. The example to be recognized is matched against the archived arrangement while considering all passable posture and scale updates [25]. Matching strategies might be gathered into three classes: immediate matching, deformable models and flexible matching and unwinding matching [32].

4.2 Syntactic or Structural Methods

In this sort of classifier an information example is grouped as far as its segments (design primitives) and the relations around them. The classifier first distinguishes the primitives of a character and after that parses series of primitives consistent with a given set of grammar tenets. Syntactic systems are generally utilized for ordering manually written content [25].

The most famous syntactic order system is to speak to characters as generation manages whose left-hand side speaks to character marks and whose right-hand side speaks to series of primitives. The right-hand side of principles is contrasted with the series of primitives concentrated from an expression. An extraordinary arrangement of order literary works has been distributed in the range of choice tree and standard based studying methods [99]. In choice trees, a character are spoken to grammatically as a tree whose inner hubs are primitives and leaves are character marks. Characterizing a character, consequently, compares to uncovering a way through the tree to a leaf. Guideline based frameworks are additionally utilized for arrangement of structural characteristics [32, 38].

4.3 Statistical Methods

Statistical classifiers comprise of mapping altered length vectors of characteristics into a parceled space [25, 32]. Arrangement here might be as straightforward as a separation classifier. Factual classifiers are immediately trainable and, when sensible suppositions are met, are moderately coldhearted to clamor. The k-NN standard is a non parametric recognition strategy. This strategy contrasts an obscure design with a set of examples that have been long ago named with class personalities in the preparation arrange. An example is distinguished to be of the class of the example to which it has the closest separation. An alternate regular measurable strategy is to utilize Bayesian order. A Bayesian classifier appoints an example to a class with the greatest a posteriori likelihood. Class models are utilized as a part of the preparation stage to gauge the class-restrictive likelihood thickness capacity for a characteristic vector. Other then these, Quadratic Discriminant Function (QDF), Linear Discriminant Function (LDF), Euclidean separation, cross association, Mahanalobis separation, Regularized Discriminant Analysis (RDA) are other factual classifiers utilized for arrangement. Covered up Markov Model (HMM) is a doubly stochastic process, with an underlying stochastic methodology that is not detectable, however might be watched through an alternate stochastic procedure that generates the grouping of perceptions. HMMs have been broadly connected to manually written word recognition [100] and corrupted content recognition [53, 54].

4.4 Artificial Neural Networks

A neural system is made out of a few layers of interconnected components called neurons. Every neuron figures a weighted total of its inputs preparing a yield indicator that is changed by means of the utilization of a direct or non-straight capacity. The fundamental point of interest of neural systems lies in the capability to be prepared immediately from illustrations, exceptional execution with boisterous information, conceivable parallel usage, and proficient instruments for studying expansive databases. Neural system methodology is non-algorithmic and is trainable. The most regularly utilized group of neural systems for example grouping assignment is the food forward system, which incorporates multilayer perceptron and Radial-Basis Function (RBF) systems [101]. A percentage of the OCR frameworks which have utilized multi-layer food forward neural systems are challenging to examine and completely fathom the choice making process. Convolutional Neural Network, Vector Quantization (VQ) systems, auto-companionship systems, Learning Vector Quantization (LVQ) are different celebrated internationally neural system strategies utilized for arrangement reason. The fundamental shortcoming of the frameworks dependent upon neural systems is their poor proficience for consensus. There is dependably a possibility of under-preparing or over preparing the framework. Plus this, a neural system does not furnish structural portrayal, which is basic from counterfeit consciousness perspective.

4.5 Kernel Methods

Kernel methods, incorporating Support Vector Machines [103], Kernel Principal Component Analysis (KPCA), Kernel Fisher Discriminant Analysis (KFDA) and so forth are gaining expanding consideration and have indicated predominant execution in example recognition. A SVM is essentially a double classifier with discriminant capacity being the weighted synthesis of piece capacities over all preparation examines. In the wake of studying by Quadratic Programming (QP), the specimens of non-zero weights are

called Support Vectors (Svs). For multi-class characterization, double Svms are joined in either one against-others or one against-one (sets astute) plan. Because of the high multifaceted nature of preparing and execution, SVM classifiers have been for the most part connected to minor classification set issues. Guaranteeing comes about have been getting for transcribed digit recognition [104]. The utilization of SVM for recognizing corrupted content is additionally expanding.

4.6 Hybrid Classifiers

All the above examined classifiers have their own focal points and burdens. Joining numerous classifiers has been as far back as anyone can remember sought after for enhancing the correctness of single classifiers. Distinctive classifiers have a tendency to differ on questionable examples, so the fusion of different classifiers can better distinguish and reject equivocal examples. Usually, consolidating correlative classifiers can enhance the grouping precision and the exchange off between slip rates and denial rate. Parallel (flat) combo is all the more regularly received for high precision, while successive (fell, vertical) synthesis is fundamentally utilized for quickening extensive classification set grouping. To enhance recognition execution, particularly for transcribed and cursive scripts, for example Arabic and Indian dialect scripts, half and half classifiers [105] have been utilized, which utilize different characteristic sorts and syntheses of classifiers masterminded in layers. It is dependent upon the thought that classifiers with diverse systems or distinctive characteristics can supplement one another. Henceforth if distinctive classifiers participate with one another, aggregate choices might decrease slips definitely and realize a higher execution. Accordingly, progressively numerous specialists now utilize blends of the above characteristic sorts and arrangement methods. Baird [106] was one of the first analysts to propose a general system for joining the qualities of structural shape examination with measurable grouping. The methodology was to build a capacity, called a characteristic distinguishing proof mapping, from the representation created by structural investigation to the one needed for factual grouping.

5. INDIAN SCRIPT RECOGNITION

As contrasted with English and Chinese dialects, the exploration on OCR of Indian dialect scripts has not realized that flawlessness. Not many endeavors have been done on the recognition of Indian character sets on Devanagari, Bangla, Tamil, Telugu, Oriya, Gurmukhi, Gujarati and Kannada. These endeavors are quickly portrayed in the accompanying sub-areas.

5.1 Devanagari and Bangla

Sinha [107-109] has begun prior finish up Devanagari script recognition. He talked about a syntactic example investigation framework and its requisition to Devanagari script recognition. Sinha and Mahabala [107] put forth a syntactic example investigation framework with an installed picture dialect for the recognition of written by hand and machine-printed Devanagari characters. The framework stores structural portrayal for every image of the script regarding primitives and their relationships. The recognition includes a quest for obscure character primitives dependent upon the archived depiction and connection. To expand the precision of the framework and diminish the computational expenses, logical data with respect to the events of certain primitives and their combos and confinements are utilized. Sinha [108, 109] later proposed information based logical post-transforming frameworks for Devanagari content recognition.

Sethi and Chatterjee [110] have additionally done some prior deal with Devanagari script. On the support of vicinity or nonappearance of some essential primitives, specifically, even line portion, vertical line section, right incline and left incline, D-bend, C-bend and so on and their positions and interconnections, they put forth a Devanagari hand-printed numeral recognition framework dependent upon binary choice tree classifier. They additionally utilized a comparative strategy for compelled hand-printed Devanagari character recognition [111]. Here, a set of extremely basic primitives is utilized, and all the Devanagari characters are looked upon as a connecting of these primitives. A multi-stage choice process is utilized where a large portion of the choices are dependent upon the vicinity / nonappearance or positional relationship of the primitives.

Connel et al. [30] have improved an online Devanagari content recognition framework. A Devanagari character recognition try different things with 20 separate journalists with every author composition five specimens of every character in a completely unconstrained manner has been led by them. An exactness of 86.5% with no rejects has been accounted for through the blend of various classifiers that concentrate on either neighborhood on-line lands, or worldwide logged off lands. Bansal [70], has advanced a learning based finish OCR for Devanagari script. Different applicable information sources have been recognized and reconciled. An execution of 70% at character level has been accomplished when the font is obscure. The execution enhanced to 80% when the font data is given to the framework. The utilization of word lexicon for amendment further upgraded the execution to 90%.

Buddy and Chaudhuri [10] have advanced an OCR framework for printed Devanagari script. Utilizing zonal data and shape characteristics, the fundamental, adjusted and compound characters are divided for the accommodation of arrangement. The adjusted and essential characters are recognized by a structural characteristic based tree classifier while the compound characters are recognized by a half and half approach, which is a consolidation of a characteristic based tree classifier and a dark run-length based glossary find. They have reported a precision of 96%.

Palit and Chaudhuri [112] have proposed a straightforward, characteristic based calculation for the PC recognition of printed Devanagari script. The calculation utilizes a binary tree-organized classifier. The classifier uses three sorts of characteristics. The beginning couple of levels are dependent upon characteristics of a consolidated run length technique. At the easier levels, nearby characteristics and minutes are utilized. A recognition rate of 90% has been accounted for.

Biswas and Chatterjee [113] have introduced a characteristic based approach to recognise Devanagari script reports utilizing a consolidation of syntactic and deterministic methodology. They have proposed two new offers from characters that help in recognizing between comparable looking Hindi characters. One of the characteristics utilizes the bearing of the ordinary at the purpose of most extreme shape of a fragment. The other characteristic is concerned with the bearing of traversal of a portion. They have tried the framework on six separate fonts and have reported a recognition rate of 96%.

After Devanagari, the following Indian dialect script on which work has been carried out is Bangla. One of the soonest endeavors for Bengali character recognition has been made by Ray and Chatterjee [114]. They exhibited a closest neighbor

classifier utilizing characteristics removed by utilizing a string connectivity paradigm for Bangla character recognition. Abusing the likeness around the major Indian scripts, Dutta [115] exhibited a summed up formal approach for era and investigation of all Bangla and Devanagari characters. Dutta and Chaudhury [116] have advanced an Isolated Optical Character Recognition (IOCR) framework for Bangla letter sets and numerals utilizing bend characteristics. Chaudhuri and Pal [76] have done noteworthy work for advancement of a complete OCR for Bangla script.

The principal complete framework equipped for doing OCR from printed Bangla records is because of Chaudhuri and Pal [76]. They have depicted an OCR framework for archives of single Bangla font [9]. The characters are differentiated into basic and compound characters and are recognized independently. The straightforward character recognition is performed utilizing a characteristic based tree classifier, and the compound character recognition includes bunching. A recognition correctness of 96% had been accounted for by the framework. Later they enhanced the Bangla OCR framework to expand the recognition precision of the framework to 99.1% for single font clear records [76]. From zonal data and shape characteristics, the fundamental, changed and compound characters are differentiated for the accommodation of grouping. The essential and changed characters are recognised by a structural-characteristic based tree classifier. The compound characters are recognised by a tree classifier emulated by arrangement matching approach. A glossary based blunder redress plan has been utilized to build the recognition correctness. Chaudhuri and Pal [117] have likewise advanced a skew discovery method for Indian dialect scripts, for example Bangla and Devanagari, which abuses the characteristic characteristics of the script to confirm the skew point. A statement in these scripts shows up as a solitary part. The upper encompass of a part is considered by section insightful checking from a fanciful line above the segment. Partitions of upper conceal fulfilling the lands of computerized straight line are identified. They are grouped as fitting in with single content lines. Gauges from distinctive bunches are joined together to get the skew plot. The system works well for distinguishing skew plots in the reach -45° to 45°.

Garain and Chaudhuri [118] proposed a technique which joins the positive parts of characteristic and run number-based standardized model matching procedures, for the recognition of printed Bangla characters. For written by hand content recognition, Pal and Datta [119] proposed a water supply based plan for the division of unconstrained written by hand message into lines, expressions and characters. Neural system approach has likewise been utilized for the recognition of Bangla characters. Dutta and Chaudhury [116] reported a take a shot at recognition of disconnected Bangla alphanumeric written by hand characters utilizing neural systems. The characters have been spoken to regarding the primitives and structural demands between the primitives encroached by the intersections show in the characters. The primitives have been characterized on the premise of the critical shape occasions like ebb and flow maxima, bend minima and inflectional focuses watched in the characters. A two stage bolster send neural net, prepared by the well-known back-proliferation calculation, has been utilized for recognition. The structural obligations encroached by the intersections have been encoded in the topology of the system itself. Bhattacharya et al. [120] have additionally utilized neural system approach for the recognition of Bangla manually written numerals. A topology adjustable self composing neural system is initially used to concentrate the skeletal shape from a numeral example. This skeletal shape is spoken to as a diagram. Certain characteristics like circles, intersections, and so on present in the diagram are acknowledged to arrange a numeral into a littler assembly. At long last, multilayer perceptron systems are utilized to characterize diverse numerals extraordinarily.

Sural and Das [121] have improved a fluffy OCR framework for Bangla script. They tried their framework on Bangla report pictures undermined by reenacted clamor and have reported a recognition rate of 98% for Bangla documents.

In the exhibited work, we have proposed new calculations, based upon the structural lands of Devnagari and Bangla script, to fragment touching characters in every one of the three zones of corrupted printed Devnagari and Bangla script [136-138]. In the center zone, different classifications of touching characters have been recognized. For fragmenting touching characters falling in first class, we distinguished the position of the sidebar and put a cut section at whatever point the sidebar sections end. Also, calculations have been created to portion touching characters for different classifications. This procedure is extremely suitable indeed, for portioning more than two touching characters at the same time in a solitary word. Comparable classifications have been characterized for upper and more level zones.

The issue of dividing evenly covering lines and partner the broken segments of a line with their separate line has been considered and comprehended in printed Devnagri and Bangla script and in seven different really popular Indian scripts [139]. An overview has additionally been distributed about the issue of division of touching characters in Indian scripts Different sorts of corruptions in debased printed Devnagri and Bangla script have been recognized as well as the issues with their recognition, wellspring of the debasements and some conceivable results have been examined for tackling the issues [141].

5.2 Tamil

The recognition of Tamil characters began in 1978 by Siromony et al. [122]. They depicted a system for recognition of machine-printed Tamil characters utilizing an encoded character string reference work. The plan utilizes string characteristics extricated by line and section astute examining of character grid. Offers in every line (section) are encoded suitably hinging on the intricacy of the script to be recognized. Chandrasekaran et al. [123] utilized comparable approach for obliged hand-printed Tamil character recognition. Chinnuswamy and Krishnamoorthy [124] introduced a methodology for hand-printed Tamil character recognition utilizing marked diagrams to depict structural arrangement of characters regarding line-like primitives. Recognition is completed by correspondence matching of the marked chart of the obscure character with that of the models.

As of late a bit of tackle line Tamil character recognition is accounted for by Aparna et al. [125]. They utilized shape based characteristics incorporating speck, line terminal, knocks and cusp. Stroke distinguishing proof is carried out by contrasting an obscure stroke and a database of strokes. Limited state mechanization has been utilized for character recognition with a precision of 71.32-91.5%.

5.3 Telugu

A two-stage recognition framework for printed Telugu letters in order has been portrayed by Rajasekaran and Deekshatulu [126]. In the first stage an administered bend following

system is utilized to recognise primitives and to concentrate essential character from the real character example. In the second stage, the essential character is coded, and on the groundwork of the learning of the primitives and the fundamental character present in the information design, the characterization is realized by method of a choice tree. Lakshmi and Patvardhan [127] displayed a Telugu OCR framework for printed content of numerous sizes and various fonts. After preprocessing, connected part methodology is utilized for division characters. True esteemed heading characteristics have been utilized for neural system based recognition framework. The creators have asserted a correctness of 98.6%. Negi et al. [128] put forth a framework for printed Telugu character recognition, utilizing joined segments and border separation based pattern matching for recognition. Border separations look at just the dark pixels and their positions between the patterns and the information pictures.

5.4 Oriya

In 1998, Mohanti [129] proposed a framework to recognise letter sets of Oriya script, utilizing kohonen neural system. The inputs pixels are sustained to the neurons in the Kohonen layer where the neurons confirm the yield consistent with a weighted aggregate equation. The character is ordered as per the biggest yield acquired from the neuron. Here the creator tried different things with just five Oriya characters and consequently the dependability of the framework is not built. In a framework advanced by Chaudhuri et al. [69] for the essential characters of Oriya script, the record picture is initially caught, preprocessed and division modules are connected. These modules have been created by joining together traditional procedures with some recently proposed ones. Next, distinct characters are recognized utilizing a blend of stroke and run-number-based characteristics, on top of characteristics acquired from their own particular planned thought of water over spill out of a repository. As of late, Roy et al. [130] have put forth a framework for disconnected from the net unconstrained Oriya written by hand numerals. They utilized histograms of bearing chain code of the shape purposes of the numerals as characteristics and a neural system based classifier has been utilized with a correctness of 94.81%.

5.5 Gurmukhi

Gurmukhi script is utilized fundamentally for composition Punjabi dialect. Punjabi Language is spoken by eighty four million local speakers and is the planet's fourteenth most broadly spoken dialect. Lehal and Singh [11, 67, 68, 77, 131] improved a complete OCR framework for printed Gurmukhi script where joined segments are initially portioned utilizing diminishing based methodology. They began work with examining handy preprocessing systems [77]. Lehal content and brings about over division or under division. The content picture is broken into and Singh [67, 68] have examined in item the division issues for Gurmukhi script. They have watched that even projection technique, which is the most regularly utilized system utilized to concentrate the lines from the record, comes up short much of the time when connected to Gurmukhi flat content strips utilizing level projection as a part of every column. The crevices on the even projection profiles are taken as separators between the content strips. Every content strip could speak to: a) Core zone of one content line comprising of upper, center zone and alternatively lower zone (center strip), b) upper zone of a content line (upper strip), c) lower zone of a content line (easier strip), d) center zone of more than one content line (multi strip). At that point utilizing assessed normal tallness of the center strip and its rate they recognize the sort of every strip. For division of strip into expressions, vertical projection is utilized and a crevice of two or more pixels in the histogram is taken to be the expression delimiter. The expression is then broken into sub-characters. To begin with, the position of the feature in the statement is considered by searching for the most prevailing column in the upper 50% of the saying. At that point, the joined segment division prepare moves ahead in three stages and at long last, each one associated part speaks to either a solitary character or a part of character lying in one of upper, center or easier zones.

In the recognition prepare, they have utilized two sorts of capabilities. In the essential list of capabilities the amount of intersections, number of circles and their positions are tried. The amount of endpoints and their area, nature of profiles of distinctive headings and so forth are acknowledged in the optional list of capabilities. A multi-stage arrangement plan consolidated with twofold tree and closest neighbor classifier has been utilized for the reason. The arrangement process is done in three stages. In the first stage, the characters are aggregated into three sets relying upon their zonal position, i.e., upper zone, center zone and easier zone. In the second stage, the characters in center zone set are further conveyed into more diminutive sub-sets by a parallel choice tree utilizing a set of powerful and font autonomous characteristics. In the third stage, the closest neighbor classifier is utilized and the exceptional characteristics recognizing the characters in every subset are utilized. One critical purpose of this plan, rather than the expected single-stage classifier where every character picture is tried against all models, is that a character picture is tried against just certain subsets of classes at every stage. This improves the computational proficiency. The framework has an exactness of something like 97.34%. An OCR post-processor of Gurmukhi script is likewise improved. In last, Lehal and Singh [132] and Lehal et al. [133] proposed a post-processor for Gurmukhi OCR where measurable data of Punjabi dialect syllable mixtures, corpora look-up and certain heuristics dependent upon Punjabi punctuation leads have been recognized.

There is likewise some expositive expression managing division of Gurmukhi Script [134, 135]. Lehal and Singh [134] have performed division of Gurmukhi script by associated segment dissection of an expression accepting the feature not being a part of the saying. Goyal et al. [135] have inferred a dismemberment based Gurmukhi character division strategy which sections the characters in the distinctive zones of a saying by analyzing the vertical white space.

For dividing touching characters, Lehal and Singh [10, 77] have utilized systems to section touching characters in Gurmukhi script in all the zones, i.e., upper, center and more level zone. For upper zone, they initially connected sectioning destination capacity proposed by Kahan et al. [52]. Likewise they broke down that if there is any eastward arranged stroke from an intersection in the second 50% of the associated part (CC), along the x-hub, which is not touching the feature. In the event that such a stroke exists, separate it from the primary part. For center zone, they utilized the procedure utilized by Kahan et al. [52] on the unthinned characters. For easier zone characters touching with one another, they have utilized the same system as utilized for upper zone.

5.6 Gujarati

Antani and Agnihotri [142] depicted grouping of a subset of printed Gujarati characters. For the characterization, least Euclidean separation and k-NN classifier were utilized with normal and invariant minutes. A Hamming separation classifier was additionally utilized.

Sets of printed Gujarati characters and modifiers were picked and subjected to characterization by Yajnik and Mohan [143] utilizing ANN architectures by recognizing direct enactment capacities in the yield layer. The specimen and test pictures for the Gujarati characters were acquired from the filtered pictures of printed Gujarati content and their characteristics were concentrated regarding wavelet coefficients. Two Multi-Layer Perceptron (MLP) arranges, one for the arrangement of letter sets which succumb to the center zone and the other one for ordering the modifiers which succumb to the easier zone are planned. These systems realize 94.46% and 96.32% precision for letters in order and modifiers, individually, on the test set.

5.7 Kannada

Ashwin and Sastry [144] reported a font and measure free OCR framework for printed Kannada reports. The proposed framework first concentrates statements from the archive picture and after that sections these into sub-character level pieces. The division calculation is roused by the structures of the script. A set of zoning characteristics is concentrated after standardization of the characters for recognition. SVM has been utilized by utilizing various two class classifiers.

An on-line framework for Kannada characters is depicted by Rao and Samuel [145]. The depicted framework separates wavelet emphasizes from the shape of the characters. The convolutional food forward multi-layer neural system is utilized as the classifier. For recognition of machine-printed Kannada script, Kumar and Ramakrishnan [146] have considered ANN based classifiers like back spread and Radial Basis Function (RBF) Networks, separated from the traditional example arrangement procedure of closest neighbor. The ANN classifiers are prepared in administered mode utilizing the convert characteristics.

Sharma et al. [147] utilized quadratic classifier based plan for the recognition of disconnected from the net manually written numerals of Kannada.

6. DEGRADED TEXT RECOGNITION

About debased content recognition a not many work has been accounted for in the writing [51, 53, 54, 148]. For Indian dialects just about nothing has been carried out on debased content recognition. Hong [51] has done work in the debased content recognition of English dialect. Consistent with him so as to enhance the execution of an OCR framework on debased pictures of content, post-handling methods are discriminating. The target of post-preparing is to redress failures or to resolution ambiguities in OCR comes about by utilizing logical data. Contingent upon the degree of connection utilized, there are distinctive levels of post-handling. In current business OCR frameworks, word level post-preparing techniques, for example concordance find, have been connected solidly. Then again, numerous OCR slips can't be revised by word level post-transforming. To conquer this impediment, section level post-preparing, in which worldwide context oriented data is used, is indispensible. In most present studies on section level post-preparing, etymological connection is the major asset to be misused. Relations at the picture level must be predictable with the relations at the typical level if word picture in the content have been translated effectively. In view of the way that OCR comes about frequently defile this consistency, systems for visual consistency investigation are intended to discover and right OCR mistakes.

A saying collocation-based unwinding calculation and a probabilistic grid parsing calculation are proposed. An intuitive model for corrupted content recognition is suggested that actualize this system as demonstrated in Figure 2.3. In this model, starting word recognition comes about are furnished by an OCR framework and they are treated as speculations to be tried further; by joining visual and semantic consistency dissection, a saying theory might be proposed, modified, denied, affirmed or chose; at long last, a choice for every expression picture is resolved.

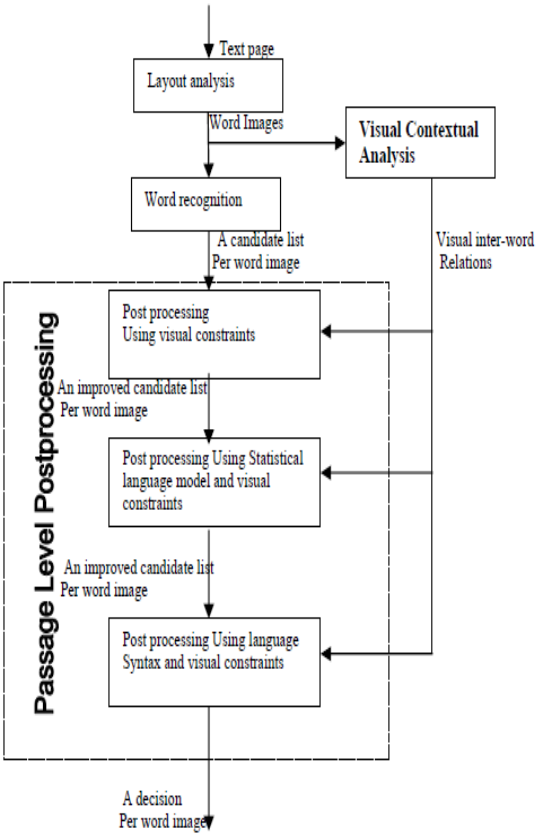


Figure 2.3: The Interactive model of text recognition system proposed by Hong [51]

Bose and Kuo [53] connected a Hidden Markov Model and level-building dynamic customizing calculation to the issues of powerful machine recognition of joined and debased characters framing expressions in a poor printed content. The recognition framework comprises of preprocessing, sub-character division and characteristic extraction, accompanied by administered studying or recognition. A structural examination calculation is utilized to section a saying into sub characters portions independent of the character limits, and to recognize the primitive characteristics in every portion, for example strokes and curves. The states of the HMM for every character are factually spoken to by the sub-character sections, and the state characteristics are gotten by verifying the state likelihood capacity, in light of the preparation inspects. Keeping in mind the end goal to recognise an obscure word, sub-character division and characteristic extraction are

performed and the move probabilities between character models are utilized for the move between characters in the string. A level-building dynamic modifying calculation consolidates division and recognition of the saying in one operation and picks the best plausible assembling of the characters for recognition of an obscure word. The machine trials exhibit the strength and viability of the new framework for recognizing expressions shaped by corrupted and associated characters.

Schenkel and Jabri [54] gathered an expansive, genuine database, holding corrupted, old and faxed records and present a correlation between two heading edge business programming bundles and human perusing execution which indicates quantitatively the colossal execution hole between people and machines, even on irregular character archives where no connection might be utilized. This demonstrates space for conceivable changes. They actualized an incorporated division and recognition calculation utilizing neural systems and HMM prepared on the database and present results, which demonstrate the prevalent execution of the calculation.

7. ACKNOWLEDGMENTS

In this paper we have present a comprehensive survey of script identification for degraded Indian language document image. Researchers are present different type of algorithm for degraded script identification for Indian language document image. In this paper we have also attempt to provide degraded different kind of script identification for Indian language document image like (Devnagari, Bangla, Tamil, telugu, kannada, Gurumukhi.

8. REFERENCES

[1] S. N. Srihari and S. W. Lam, "Character recognition", Center of Excellence for Document Analysis and Recognition (CEDAR), *Technical Report*, 1995.

[2] M. E. Stevens, "Automatic character recognition-State-of-the-art report", National Bureau of Standards & Technology, Tech. Note 112, Washington, USA, 1961.

[3] S. Mori, C. Y. Suen and K. Yamamoto, "Historical review of OCR research and development", *Proceedings of the IEEE*, Vol. 80(7), pp. 1029-1058, 1992.

[4] U. Pal and B. B. Chaudhuri, "Indian script character recognition: a survey", *Pattern Recognition*, Vol. 37(9), pp. 1887-1899, 2004.

[5] S. Mori, K. Yamamoto and M. Yasuda, "Research on machine recognition of handprinted characters", *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, Vol. 6(4), pp. 386–405, 1984.

[6] G. Nagy, "At the frontiers of OCR", *Proceedings of the IEEE*, Vol. 80(7), pp. 1093-1100, 1992.

[7] R. Plamondon and S. N. Srihari, "On-line and off-line handwritten recognition: a comprehensive survey", *IEEE Transactions on PAMI*, Vol. 22(1), pp. 63–84, 2000.

[8] C. Y. Suen, R. Legault, C. Nadal, M. Cheriet and L. Lam, "Building a new generation of handwriting recognition systems", *Pattern Recognition Letters*, Vol. 14(4), pp.305-315, 1993.

[9] U. Pal and B. B. Chaudhuri, "Computer recognition of printed Bangla script",*International Journal of Systems Science*, Vol. 26, pp. 2107-2123, 1995.

[10] U. Pal and B. B. Chaudhuri, "Printed Devanagari script OCR system", *Vivek*, Vol.10(1), pp. 12-24, 1997.

[11] G. S. Lehal and C. Singh, "A complete machine printed Gurmukhi OCR system", *Vivek*, Vol. 16(3), pp. 10-17, 2006.

[12] M. Bosker, "Omnidocument technologies", *Proceedings of the IEEE*, Vol. 80(7), pp.1066-1078, July 1992.

[13] T. Pavlidis, "Problems in recognising poorly printed text", in Symposium on Document Analysis and Information Retrieval(SDAIR), pp. 163-172, 1992.

[14] S. V. Rice, F. R. Jenkins and T. A. Nartker, "The fourth annual test of OCR accuracy", Technical Report 95-04, ISRI, University of Nevada, Las Vegas, pp. 11-50, 1995.

[15] J. Rocha and T. Pavlidis, "A solution to the problem of touching and broken characters", in the Proceedings of the 2nd International Conference on Document Analysis and Recognition (ICDAR), pp. 602-605, 1993.

[16] C. Fang, *Deciphering Algorithms for Degraded Document Recognition*, Ph. D. thesis, State University of New York at Buffalo, 1997.

[17] H. Bunke and P. S. P. Wang, *Handbook of Character Recognition and Document Image Analysis*, World Scientific Publishing Company, 1997.

[18] Stephen V. Rice, George Nagy and Thomas A. Nartker, *Optical Character Recognition: An Illustrated Guide to the Frontier*, Kluwer Academic Publications, 1999.

[19] S. Mori, H. Nishida and H. Yamada, *Optical Character Recognition*, John Wiley & Sons, 1999.

[20] J. Mantas, "An overview of character recognition methodologies", *Pattern Recognition*, Vol. 19(6), pp. 425-430, 1986.

[21] V. K. Govindan and A. P. Shivaprasad, "Character recognition-A review", *Pattern Recognition*, Vol. 23 (7), pp. 671-683, 1990.

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

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

[24] Q. Tian. P. Zhang. T. Alexander and Y. Kim, "Survey: omnifont-printed character recognition", in the proceedings of Visual Communications and Image Processing SPIE, Vol. 1606, pp. 260-268, 1991.

[25] A. K. Jain, R. P. W. Duin and J. Mao, "Statistical pattern recognition: a review", *IEEE Transactions on PAMI*, Vol. 22(1), pp. 4-37, 2000.

[26] R. Kasturi and L. O'Gorman, "Document image analysis: a bibliography", *Machine Vision and Applications*, Vol. 5(3), pp. 231-243, 1992.

[27] C. C Tappert, C. Y. Suen and T. Wakahara, "The state of the art in on-line handwriting recognition", *IEEE Transactions on PAMI*, Vol. 12(8), pp. 787-808,1990.

[28] T. Wakahara, H. Murase and K. Odaka, "On-line handwriting recognition",*Proceedings of the IEEE*, Vol. 80(7), pp. 1181-1194, 1992.

[29] F. Nouboud and R. Plamondon, "On-line recognition handprinted characters: survey and beta tests", *Pattern Recognition*, Vol. 23(9), pp. 1031-1044, 1990.

[30] S. D. Connell, R. M. K. Sinha and A. K. Jain, "Recognition of unconstrained **o**n-line Devanagari characters", in the Proceedings of 15th International Conference on Pattern Recognition (ICPR), Vol. 2, Spain, pp. 368-371, 2000.

31] S. D. Connell and A. K. Jain, "Template-based online character recognition", *Pattern Recognition*, Vol. 34(1), pp. 1-14, 2001.

[32] F. Bortolozzi, A. Britto Jr., L. S.Oliveria and M. Morita, "Recent advances in handwriting recognition", in the Proceedings of International Workshop on Document Analysis (IWDA), India, pp. 1-30, 2005.

[33] S. W. Lee, "Off-line recognition of totally unconstrained handwritten numerals Using multiplayer cluster neural network", *IEEE Transactions on PAMI*, Vol. 18(6), pp.648-652, 1996.

[34] F. El-Khaly and M. A. Sid-Ahmed, "Machine recognition of optically captured machine printed Arabic text", *Pattern Recognition*, Vol. 23(11), pp. 1207-1214, 1990.

[35] A. Amin, "Off-line Arabic character recognition- a survey", in the Proceedings of 4th ICDAR, pp. 596-599, 1997.

[36] L. M. Lorigo and V. Govindaraju, "Offline Arabic handwriting recognition: a survey", *IEEE Transactions on PAMI*, Vol. 28(5), pp. 712-724, 2006.

[37] T. H. Hildebrandt and W. Liu, "Optical recognition of handwritten Chinese characters: Advances since 1980", *Pattern Recognition*, Vol. 26(2), pp. 205-225,1993.

[38] C. L. Liu, S. Jaeger and Masaki Nakagawa, "Online recognition of Chinese characters: the state-of-the-art", *IEEE Transactions on PAMI*, Vol. 26(2), pp. 198- 213, 2004.

[39] R. G. Casey and E. Lecolinet, "A survey of methods and strategies in character segmentation", *IEEE Transactions on PAMI*, Vol. 18(7), pp. 690-706, 1996.

[40] C. E. Dunn and P. S. P. Wang, "Character segmentation techniques for handwritten text - a survey", in the Proceedings of 11th ICPR, Vol. 2, pp. 577-580, 1992.

[41] Y. Lu, "Machine printed character segmentation - an overview", *Pattern Recognition*,Vol. 28(1), pp. 67-80, 1995.

[42] Y. Lu and M. Shridhar, "Character segmentation in handwritten words – an overview", *Pattern Recognition*, Vol. 29(1), pp. 77-96, 1996.

[43] R. L. Hoffman and J. W. McCullough, "Segmentation methods for recognition of machine-printed characters", *IBM Journal of Research and Development*, Vol. 15(2),pp. 153-165, 1971.

[44] H. S. Baird, S. Kahan and T. Pavlidis, "Components of an omnifont page reader", in the Proceedings of 8th ICPR, Paris, pp. 344-348, 1986.

[45] S. Tsujimoto and H. Asada, 1992, "Major components of a complete text reading system", *Proceedings of the IEEE*, Vol. 80(7), pp. 1133-1149, 1992.

[46] S. Liang, M. Shridhar and M. Ahmadi, "Segmentation of touching characters in printed document recognition", *Pattern Recognition*, Vol. 27(6), pp. 825-840, 1994.

[47] J. Wang and J. S. N. Jean, "Segmentation of merged characters by neural networks and shortest path", *Pattern Recognition*, Vol. 27(5), pp. 649-658, 1994.

[48] R. G. Casey and G. Nagy, "Recursive segmentation and classification of composite character patterns", in the Proceedings of 6th ICPR, pp. 1023-1026, 1982.

[49] S. W. Lee, D. J. Lee and H. S. Park, "A new methodology for gray-scale character segmentation and recognition", *IEEE Transactions on PAMI*, Vol. 18(10), pp. 1045- 1050, 1996.

[50] S. Tsujimoto and H. Asada, "Resolving ambiguity in segmenting touching characters", in the Proceedings of 1st ICDAR, pp. 701-709, 1991.