Feature Extraction of Degraded Devanagari Joint Words And Numerals Using Zernik Moment Feature extraction and Hybrid Feature extraction Methods for Recognition and conversion into editable form

Sushilkumar N. Holambe[1],Ulhas B Shinde[2]
[1]Computer Science & Engineering Department, College Of Engineering, Osmanabad-413512(M.S.)India.
[2]CSMSS,CSCOE, Aurangabad-431001(M.S.) India
e-mail:-snholambe2015@gamil.com

Abstract
The feature extraction technique is projected in favor of corrupted Devnagari characters scan manuscript by bearing in mind the arrangement of Devnagari characters .So we encompass and too think about the abnormality of Devnagari lettering technique, Yuktaksha as well as figures. We be not eliminate shirorekha. It take apart the picture transcript credentials into outline, terms and typescript Segmentation of tetchy or merged typescript of Devanagari characters is occasionally not easy due to interline gap or partly cover as well as blare. The algorithmic regulation worn at this time is meant for segmentation of join together Devanagari typescript addicted to its essential partial or complete consonants. Within our planned method, picture Binarzation meant for corrupted manuscript imagery have being employ area support segmentation. Initially, an RGB picture translate keen on gray picture after that picture strain be able to be completed on the source of Wiener filter and Gaussian filter..

Keywords
Devanagari lettering, Conjunct Script, Ocular Character Identification, Word Segmentation, Feature Extraction, Zernike moment, Hybrid Feature

1. Introduction
Ocular Character Identification is that the method of translating pictures of written, on paper transcript into a arrangement understood by equipment. The worlds in order of journalism, olden times, in addition to additional information be in hard-copy credentials. OCR method translate this in order by exchange the transcript on manuscript keen on electronic shape. Ocular Character Identification method is resourcefully urbanized for character detection of Non-Indian speech, since the complication of font is a minor quantity since judge against to Indian characters.
Resourceful Indian verbal communication OCR essentially depends ahead the preprocessing stage (word segmentation) pro improved identification of compound or conjunct font. Consequently, in general achievement velocity linked correctness of an Indian lettering OCR method depends ahead on the accurate segmentation of typescript.
It be necessary to verge the article picture consistently in arrange to take out valuable data in addition to create additional dispensation such while font identification and characteristic extraction, particularly for those deprived class manuscript figures with darkness, non-uniform enlightenment, little disparity, huge indication reliant clamor, wipe and blur..
The techniques of segmentation be generally classify as follows:

- Classical approach: During this method the segmentation be recognized through take out the distinctive feature of the font picture.
- Recognition based segmentation: During this method the picture as a entire is investigate pro apparatus so as to equal predefined module.
- Holistic approach: The methods seek to identify the expression the same as a entire.

Presently there are on the subject of fifty essential font in lettering. Within a expression, the vowel typescript frequently get modified form describe as modifiers. Consonant modifiers be too likely. The essential and complex typescript is linked through modifiers toward obtain novel form. Separately as of these the papers printed throughout this lettering demonstrate huge dissimilarity inside typescript features, variety drawing, and in font sizes. So line name shirorekha within Devanagari along with referred since caption. The adjacent font of a character extremely frequently feel from first to last the caption to shape a linked module.OCR be the most demanding area of explore in characters sample identification, text handing out, and picture dispensation. Devanagari writing be the the majority of ordinary speech recognized as Hindi within Indian province. The Devanagari lettering characteristic taking out have be a brave pro Hindi lettering OCR.

2. Literature Review

So many technology are planned through more than a few investigators on picture segmentation by means of binarization in addition to its purpose in the direction of touching entity recognition plus individual pace identification. A assessment of the fresh investigate taking place as binarization be known as.

Sauvola et al (2000) [HYPERLINK "Jaa00" 9]obtainable a fresh technique intended for adaptive text picture binarization, somewhere the sheet be measured like a anthology of subconstituents such like manuscript, surroundings as well as image. Assessment of general threshold be support on evaluation of general signify as well as home customary variation.

Randolph et al (2001) 10]recommended a dual area move toward so as to improve fax credentials via directional strain depository allow boundaries in addition to curve within the manuscript writing near live sharp suitably. It be worn in favor of better in Fax papers. A directional strain depository contain be worn so as to is competent for soft of boundaries as well as curve.

Wu et al (2003) [HYPERLINK \l "Adn03" 11]research through a multi-phase worldwide thresholding move toward go behind via a general spatial thresholding, the mechanism fine for easy as well as multifaceted imagery of postal encloses. During primary phase worldwide thresholding method worn. During next phase modification of doorsill value is completed.

Arora and Sandhya [8] offered plus OCR in support of printed Devnagari font. Essential secret code be documented through neural classifier in addition to worn four characteristic taking out method specifically, meeting point, gloominess characteristic, sequence regulations histogram with directly line appropriate characteristics. Kompalli and Suryaprakash considered OCR of Devanagari writing at hand a broad variety of confront so as to not observe within Latin support lettering. Yogesh Dandawate et. al. projected a technique inside which the typescript be scrutinize, preprocessed as well as on each human being typescript wavelet convert be functional consequently since toward obtain decaying imagery of font. The correctness acquire be approximately 85 percent. Kumar et al. projected a Zernike instant foundation
characteristic withdrawal as well as used artificial neural system classifier.

3.DEVNAGRAI SCRIPT

Lettering technique inside the characters be straight, left toward right as well as too the typescript don't contain several highercase/subordinatecase dissimilarity. Presently be concerning fifty fundamental lettering within script encompass almost single toward single association (the majority of them be still identify via the similar forename). Inside a expression, a vowel typescript typically get distorted form submitred toward like modifiers. The Consonant modifiers be as well achievable. In addition, among two and four consonants be able to join toward shape close to concerning 250 complex characters, so that partially preserve the form of the ingredient consonants. In accumulation, the majority of the primary plus complex fonts be frequently enthusiastic awake through modifiers toward acquire fresh form. Separately as of these, the credentials on paper inside these lettering demonstrate huge difference inside typeset faces, kind plan, plus inside character dimensions. On behalf of an huge diversity of typescript it's leaving toward be illustrious so as to present survive a straight line on the superior semi. The line is describe as shirorekha within Devnagari as well as be referred at this juncture like caption. The adjacent typescript of a expression quite frequently small piece from side to side the caption near generate a associated ingredient. Within lettering, a manuscript expression be too separation off keen on 3 region. The superior region indicate the piece on zenith of the caption (ascenders), the middle region wrap the the majority piece of the necessary plus composite typescript as well as in addition the subordinate region, so a number of vowel as well as consonant modifiers be able to exist. The Devnagari vowels be not sprinkled within the 'Varnamala' except be set next to the start alphabet.

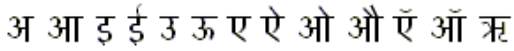

Figure 1.

The Devnagari Vyanjans

They be incredibly sensibly set into subsequent set.

A.The Sparsh
B.The Antashth
C.The Ushm

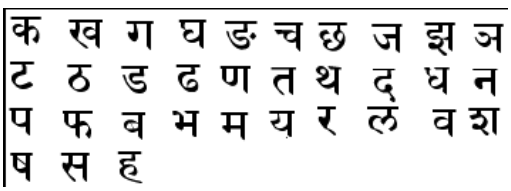

Figure 2.

Here in the subsequent fig we subsist to observe swars (vowels as well as diphthongs) plus their numerous matras. The important scattered round stand for a placeholders on behalf of consonants (or foundation-letters), therefore we resolve distinguish the comparative location of folks matras.

अ	आ	इ	ई	उ	ऊ
	ा	ि	ी	ु	ू

ए	ऐ	ओ	औ	ऍ	ऑ	ऋ
े	ै	ो	ौ	ॅ	ॉ	ृ

Figure 3.

I.The shuddh _vyaNjan (semi shape)

The shuddh vyaNjan means uncontaminated consonants. These four-sided figure calculate folks vyaNjan so as to four-sided figure calculate marked at the same time since not the intrinsic vowel. So while converting the vyaNjan we didn't mark sprawling. At the same time as typewriting syllabify we contain a propensity toward obtain these unpolluted consonants through typescript haling one time the usual variety. Here the aakaar (perpendicular line) be frequently detached following we mark the shuddh_ vyaNjan..

क् ख् ग् घ् ङ्

च् छ् द् ध् न्

श् ष् स् ह् ळ्

Figure 4.

II.The Nuktaa _vyaNjan

The Nuktaa be a accent blot. Subsequent are
ड़
ढ़
vyaNjan be awfully ordinary inside Hindi.

III.The chandra-bindu

The Chandra-bindu be a nasal mark.

With ि, ी, े, ै, ो and ौ instead of ं (chandra-bindu) people generally write ं (Anusvaar); i.e. instead of किं, कीं, कें, कैं, कों and कौं people generally write किं, कीं, कें, कैं, कों and कौं.

IV.The joDda_AkShar

The joDda is describe as link or combined. Consequently united script (conjuncts) be describe as the joDda_AkShar.

क्ष त्र ज्ञ श्र

Figure 5.

V.The Devnagari Numerals

० १ २ ३ ४ ५ ६ ७ ८ ९

4.THE FEATURE EXTRACTION

INSTANT FUNCTIONS

Instant carry out four-sided figure compute draw round on figure since the prejudiced arithmetic of the representation strength purpose. Moment purpose of arrange (p+q) be characteristically sketch as

$$\phi_{pq} = \int_x \int_y \psi_{pq}(x,y)\, f(x,y)\, dxdy,$$

here $\psi_{pq}(x,y)$ be describe the moment weighting most important part.

When the be relevant moment purpose to digital images it's typically enthralling to inscribe downward them elsewhere maltreatment the following dissimilar memorandum:

$$\phi_{pq} = \sum_x \sum_y \psi_{pq}(x,y)\, f(x,y).$$

a number of belongings of the massing most important part are approved on the instant themselves, such since inconsistency features. Depending on top of the execute choose in favor of the coefficient most important part, the intended instant determination imprison completely dissimilar feature of the input picture[24].

5. Zernike Moments.

Since opposition arithmetical instant, Zernike Moments four-sided figure calculate sketch in excess of the component quite than diagram as well as show the orthogonality possessions. The Zernike polynomials be mostly worn inside optometric, anywhere they happen since the growth of a gesticulate frontage function inside visual method by means of spherical pupil [5]. The Zernike bring in a compilation of higher polynomials so as to variety of whole orthogonal situate in excess of the within of the element ring, i.e., $x^2 + y^2 = 1$. allow the place of these polynomials exist indicate by means of $\{V_{nm}(x,y)\}$. The shape of these polynomials be:

$$V_{nm}(x,y) = V_{nm}(\rho\ \theta) = V_{nm}(\rho)\exp(jm\theta)$$

$R_{nm}(\rho)$ The Radial Polynomial of distinct as

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

reminder that $R_{n.-m}(\rho) = R_{nm}(\rho).$

These polynomials be orthogonal as well as make happy

$$\int\int_{x^2+y^2\le 1} \left[V_{nm}(x,y)\right]^{*} V_{pq}(x,y)\, dx\, dy = \frac{\pi}{n+1}\delta_{np}\delta_{mq}$$

with $\delta ab = \begin{cases} 1 & a=b \\ 0 & otherwise. \end{cases}$ Zernike moments four-sided figure measure the outcrop of the picture take out on these orthogonal foundation purpose. The zernike moment of arrange n by means of reincidence m in favor of

$$A_{nm} = \frac{n+1}{\pi}\int\int_{x^2+y^2\le 1} f(x,y)\ V_{nm}^{*}(\rho,\theta)\, dxdy$$

in favor of a digital picture, the integrals be put back through summations in the direction of obtain.

$$A_{nm} = \frac{n+1}{\pi}\sum_x \sum_y f(x,y)\ V_{nm}^{*}(\rho,\theta),\ \ x^2 + y^2 \le 1.$$

in the direction of compute the zernike moments of a known picture, the middle of the picture be in use since the source in addition to pixel organize be plan in the direction of the variety of component ring, i.e. $x^2 + y^2 \le 1.$ folks pixels declining

exterior the component sphere don't appear in the direction of exist working inside the calculation. As well reminder that $A^{*}_{nm}\, A_{n.-m}$. consequently |Anm| be able to be worn as a turning round invariant characteristic of the picture purpose.

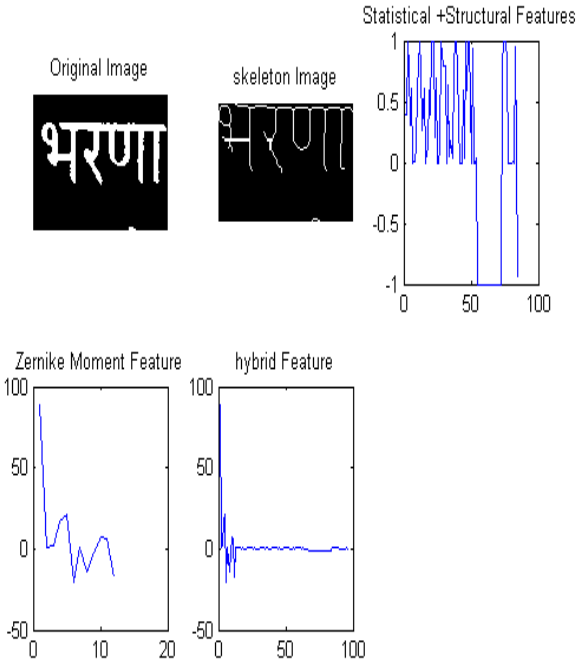

Fig 6:Feature extraction of degraded devanagari word

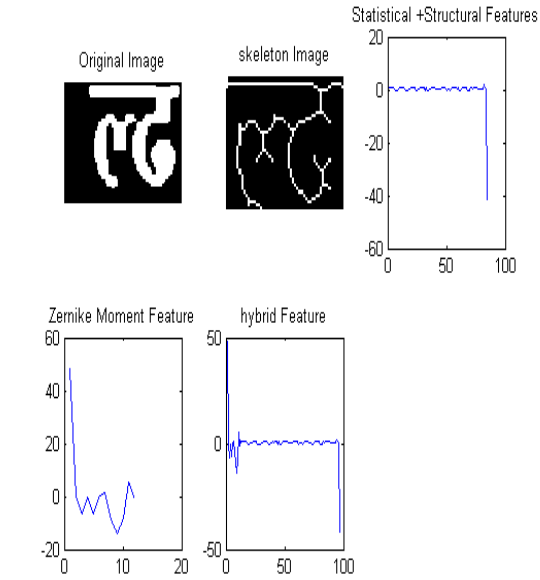

Fig 7:Feature extraction of degraded devanagari Joint Word

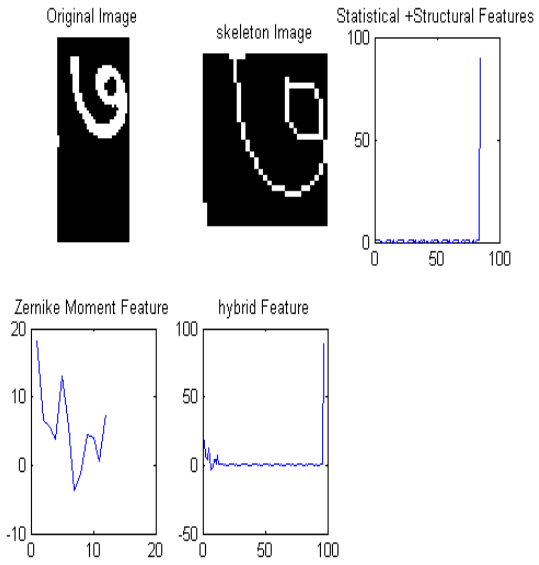

Fig 8:Feature extraction of degraded devanagari numeral.

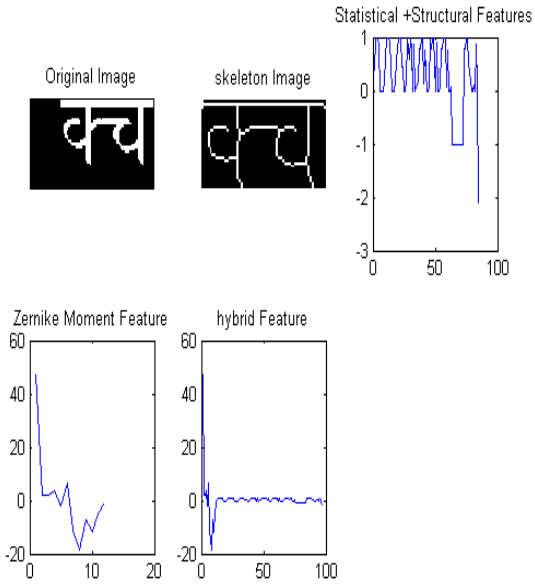

Fig 9:Feature extraction of degraded devanagari joint word.

6.The Zernik Moment Feature drawing out and Hybrid Feature drawing out.

Here in the Zernik Moment Feature drawing out as well as Hybrid Feature drawing out, preparation prototype is plan inside d-dimensional breathing space, someplace d be the numeral of characteristics there. So these prototype four-sided shape calculate designed reliable by means of their exposed characteristic standards as well as four-sided figure calculate label reliable by means of their improved recognized group. An untaged examination prototype be plan inside the identical room in addition to be confidential concurrenceing in the direction of the majority of regularly happening group in the middle of its K-most alike preparation prototype to its adjacent neighbors. Therefore mainly ordinary resemblance compute on behalf of knn categorization be the Euclidian space metric, distinct flanked by characteristic vectors $\vec{x}$ *and* $\vec{y}$ as :

$$euc(\vec{x},\vec{y}) = \sqrt{\sum_{i=1}^{f}(x_i - y_i)^2}$$

here f stand for the figure of characteristic. Slighter space worth stand for better resemblance.

7.Results

TABLE II.RECOGNITION FEATURE EXTRACTION IN %

TABLE I. ERROR RATE IN %

	ZMF	HF
DS1	8	5
DS2	5	4
DS3	6	7
DS4	8	8
DS5	7	8
DS6	7	5

	ZMF	HF
DS1	95	96
DS2	96	95
DS3	97	94
DS4	87	91
DS5	87	92
DS6	98	97

So here we can see that by means of these two feature extraction methods we can extracte feature of every devanagari dataset and we can perform the classification.The results are given above.Here we not separated numeral,symbols,vowels and consonant. We have combined them.These feature extraction methods are very important in recognition of degraded devnagari script scan documents.

6.ACKNOWLEDGMENT

I am very much grateful to all my senior friends and unknown people who have helped in creation of dataset which is important in recognition of degraded devnagari script scan documents..

7.REFERENCES

[1] V. Bansal and R. M. K. Sinha, "Integrating nowledge Sources in Devnagri Text Recognition," IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans, vol. 30, pp. 500 – 505, July 2000.

[2] Meng Shi, Yoshiharu Fujisawa, Tetsushi Wakabayashi, and Fumitaka Kimura. ,""Handwritten numeral recognition using gradient and curvature of gray scale image." Pattern Recognition, 35(10):2051−2059,2002.

[3] Liu Cheng-Lin, Nakashima, Kazuki, H,Sako, and H.Fujisawa, "Handwritten digit recognition: investigation of normalization and feature extraction techniques", Pattern Recognition, Vol. 37, No. 2, pp. 265-279 , 2004.

[4] T.M. Cover and P. E. Hart, "Nearest neighbor pattern classification," IEEE Trans. Inform. Theory, vol. IT-13, pp. 21-27, Jan. 1967.

[5] V. Bansal, R.M.K. Sinha, "Partitioning and searching dictionary for correction of optically read Devnagari characters strings", Proceedings of the Fifth International Conference on Document Analysis and Recognition , 1999, pp. 653-656.

[6] S. Arora, D.Bhattacharya, M. Nasipuri, L.Malik, "A Novel Approach for Handwritten Devanagari CharacterRecognition" in IEEE –International Conference on Signal And Image Processing, Hubli, Karnataka, Dec 7-9, 2006

[7] Adawait Dixit, Ashwini Navghane, Yogesh Dandawate, "Handwritten Devanagari Character Recognition using Wavelet Based Feature Extraction and Classification Scheme", 11 th IEEE India conference INDICON, ISBN no.978-1-4799-5362-2,2014.

[8] Satish Kumar and Chandan Singh,"A Study of Zernike moments and it's use in devanagri Handwritten character recognition "Int. Conf. on Recognition, pp.514-520, 2005.

[9] Recognition of handwritten Devnagri Numerals ",In Proc. of the workshop on learning algorithm for pattern recognition , Sydney, pp.1-7,2005.

[10] N.Sharma,U.Pal,F.Kimura and S.Pal,"Recognition of offline handwritten devanagri characters using quadratic classifiers", ICVGIP, LNCS4338, pp.805-816 ,2006.

[11] Bikash Shaw, Swapan Kumar Parui, Malayappan Shridhar, "Offline Handwritten Devanagari Word Recognition: A Holistic Approach Based on Directional Chain Code

[12] G.S. Lehal and Nivedan Bhatt, " A recognition system for Devnagri and English handwritten numerals", Advances in Multimodal Interfaces– ICMI 2001, T. Tan, Y. Shi and W. Gao (Editors), LNCS, Vol. 1948, 2000, pp. 442-449.

[13] F.Kimura, Y.Miyake, M.Shridhar,"Relationship Among Quadratic Discriminant Functions for Pattern Recognition",Proc. Of 4Th IWFHR(1994).

[14] T.M. Cover and P. E. Hart, "Nearest neighbor pattern classification," IEEE Trans. Inform. Theory, vol. IT-13, pp. 21-27, Jan. 1967.

[15] B. V. Dasarathy, Nearest neighbor pattern classification techniques. IEEE Computer Society Press, New York, 1991.

[16] V. Bansal and R. M. K. Sinha, "Integrating nowledge Sources in Devnagri Text Recognition," IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans, vol. 30, pp. 500 – 505, July 2000.

[17] C.J.C. Burges. A tutorial on support vector machines. Data Mining and Knowledge Discovery, 2, 1998.

[18] Rafael C. Gonzalez and Richard E. Woods. Digital Image Processing, Second Edition. Prentice Hall,2002.

[19] Liu Cheng-Lin, Nakashima, Kazuki, H,Sako, and H.Fujisawa, "Handwritten digit recognition: investigation of normalization and feature extraction techniques", Pattern Recognition, Vol. 37, No. 2, pp. 265-279 , 2004. .

[20] J. Cao, M. Ahmadi, and M. Shridhar, "Handwritten numeral recognition with multiple features and multistage classifiers," in IEEE International Symposium on Circuits and Systems, vol. 6, (London), pp. 323-326, May 30-June 2 1994.

[21] S. Perantonis Gatos B I. Pratikakis, "An Adaptive Binarization Technique for Low Quality Historical Documents," in Document Analysis Systems VI, vol. 3163, , 2004.

[22] Chen Y. and G. Leedham, "Decompose algorithm for thresholding degraded historical document images," in IEEE Proc.-Vis. Image Signal Process vol. 152, , December 2005.

[23] Chen Y. and G. Leedham, "Document binarization using Kohonen," in IET Image Process, 2007, pp. 67-85.

[24] Ioannis Pratikakis, and Stavros J. Perantonis. Gatos Basilios, "Improved document image binarization by using a combination of multiple binarization techniques and adapted edge information," in Pattern Recognition, vol. ICPR 2008. 19th International Conference on. IEEE, 2008, 2008.

[25] Konstantinos Ntirogiannis, and Ioannis Pratikakis Gatos Basilios, "ICDAR 2009 Document Image Binarization Contest (DIBCO 2009)," in ICDAR, vol. 9, 2009.