A language independent Characterization of Document Image Noise in Historical Scripts

Sandhya.N
Dept. of Computer Science & Engineering
Dayananda Sagar College of Engineering, Bangalore, India

R. Krishnan
Dept. of Computer Science & Engineering
Dayananda Sagar College of Engineering, Bangalore, India

D. R. Ramesh Babu
Dept. of Computer Science & Engineering
Dayananda Sagar College of Engineering, Bangalore, India

ABSTRACT
Digitization of historical documents helps preserve these documents. As these documents have existed for a long time, various types of noise creep in. In our paper we have analyzed the different types of noise that occur in printed and handwritten historical documents mainly based on Kannada (Kannada is a language used in Karnataka, a southern state in India) documents and created a taxonomy for the same. We have also characterized each noise type based on factors such as their source, their effect on characters and the associated challenges in character recognition. We have also catalogued the different noise detection, removal and restoration techniques that are reported in the literature for each of the prominent noise types, and identified areas relating to noise detection, removal for further research focus.

Keywords
optical character recognition, global noise, local noise.

1. INTRODUCTION
Optical Character Recognition (OCR) is an active research area in the field of Document Image Analysis. Many OCR systems for different languages exist today which recognize clear printed documents. But handling degraded characters is a major challenge for OCR systems [19]. In practice there is need to recognize characters affected by noise in documents. Noise like skew, shadow, blur, bleed through, paper texture etc. affect the characters contained in a document and make them degraded. The noise may be introduced due to scanning, photocopying or camera captured image. In historical documents ageing is a major factor that introduces a lot of noise. These noise types lead to poor quality image of the document. It is important to improve the quality of the document by removing the noise caused by these frequently occurring problems, as they may cause the OCR system to fail. Fig 1 shows the context diagram for a typical OCR system envisaged for preserving historical scripts.

The digitization of documents is of great demand in libraries, archives etc in order to conserve resources and also to preserve the old documents. The digitization is done by scanning the present documents and preserving them in hard disks, uploading them to the web sites to provide online information. While scanning these documents, noise will be introduced due to the procedure of scanning or due to noise present in the document itself. But as part of digitization the noise has to be removed. Noise in the documents leads to degradation of characters. Fig 2a and 2b shows character degradation due to bleed through noise and shows the same character without degradation for reference. Recognition of such degraded characters is a challenging research problem and is of great value in applications like digitization and archiving of historic documents.

There are varieties of noise types. Removal techniques for one noise type will not work for other types. Therefore most of the researchers work in depth with each type of noise and find a novel idea for each. There is a lot of scope for many noise types where no work is reported as given in [5].

As noise is a key factor that affects the efficiency of OCR systems, we have analyzed the occurrence of different types of noise in historic scripts and identified the top five based on frequency of occurrence. For these prominent noise types we have catalogued the detection, removal and restoration techniques.

In section 2 we present a taxonomy of noise that occur in historical documents. Section 3 details the prominent noise types like skew, border/marginal, noisy background, deformed/degraded characters, and touching characters their detection, removal and restoration techniques. In conclusion we discuss the challenges in this area and our plans for future work.

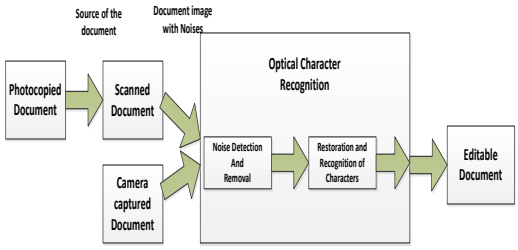

Fig 1: Context diagram of OCR system

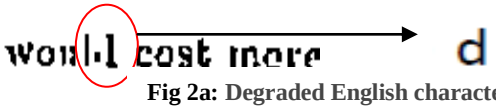

Fig 2a: Degraded English character and the character without degradation

Fig 2b: Degraded Kannada character and the character without degradation

2. TAXONOMY OF NOISE

The different possible types of noise in a document image are listed in Table 1. These noise types can be broadly classified into global and local noise. Global noise is one which manifests itself throughout the document like skew, salt and pepper noise, dark spots etc. Fig 3 shows a document which has a global noise, in this case a noisy background. The noise is spread throughout the document. Local noise is typically content oriented noise such as broken character, degraded character etc. Fig 4 is a document which has local noise, in this case reader's annotations is in portions of the document. During the recognition of the characters, these annotations have to be omitted. Fig 5 shows the taxonomy in which the noise listed in Table 1 is categorized. In this taxonomy global and local noise are further classified based on the factors which cause them. The main factors which cause noise are: digitization and storage, ageing, physical factors and document related factors.

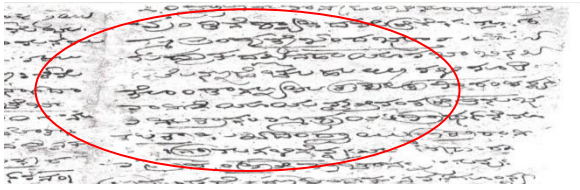

Fig 3: Document with Global noise in Kannada document

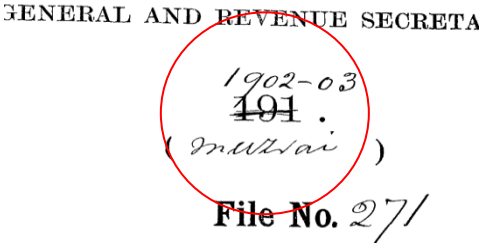

Fig 4: Document with local noise in English document

2.1 Ageing

All historical documents suffer from noise due to ageing factor. As the papers become old the color of the paper changes which introduces background noise. The paper becomes brittle and powdered which introduces noises such as worm holes, scratches and cracks, and noisy background (due to yellowing and fading of paper). Fig 6 shows examples of noise from ageing with the noisy part circled, for focus. The noisy background can be clearly seen throughout the document.

2.2 Digitization and Storage

Digitization is scanning/photocopying documents to films rolls or to disks in standard digital formats for the purpose of storage. During this process, noise types are introduced such as skew, multi-skew (each line of the document has different skew angle), warping, shadowing, noisy background, degraded characters, marginal/border noise. Fig 7 shows samples of noise introduced during digitization.

2.3 Physical factors

Some of the physical factors which create noise are the carbon copy, paper texture, folding marks, stains, sunburn (over exposure to sunlight) , thorn-off region (e.g.: papers separated from stapled document) and translucent papers. Fig 8 shows samples of physical noise.

Table 1. Possible noise types

1.Skew, Multi-skew	**2.Varying fonts**	**3.Mixed-alphabets**	**4.Shadow effect**	**5.Warping**	**6.Noisy background**	**7.Sunburn**	**8.Paper punching marks**
9.Scratches and cracks	**10.Marginal noise/border**	**11.Stains**	**12.Worm holes**	**13.Blurring**	**14.Carbon copy effect**	**15.Salt and pepper noise**	**16.Folding marks**
17.Thorn-off regions	**18.Paper texture**	**19.Bleed-through**	**20.Corrections/ additions**	**21.Touching characters**	**22.Deformed or degraded characters**	**23.Reader's annotations, seal, stamp**	**24.Inadequate printing**

2.4 Document related factors

The noise which is present in the document content due to its relation to the document are document related factors. For example in Fig 4 Reader's annotation may be comments, signatures, seal etc. is the noise related to the document, but it poses problems for OCR systems to recognize them since usually they will be of different orientation and shapes. Others are corrections/additions, mixed alphabets, touching are characters and varying fonts which are local noise types. Fig 9 shows samples of document related noise.

3. Prominent noise types

We collected a dataset of 4000 document images which are historical Kannada scripts (both printed and handwritten), from the Karnataka state archives department. We analyzed this dataset to identify the top five prominent noise types. These 4000 samples were split into four different subsets based on chronology namely: dataset1, dataset2, dataset3 and dataset4.

- Dataset 1 is handwritten Kannada document of the year 1670 to 1827 which is highly degraded.
- Dataset 2 is handwritten Kannada document of the year 1834 to 1874.
- Dataset 3 is printed Kannada document of the year 1895 to 1974.
- Dataset 4 is printed Kannada document of 1911 to 1920.

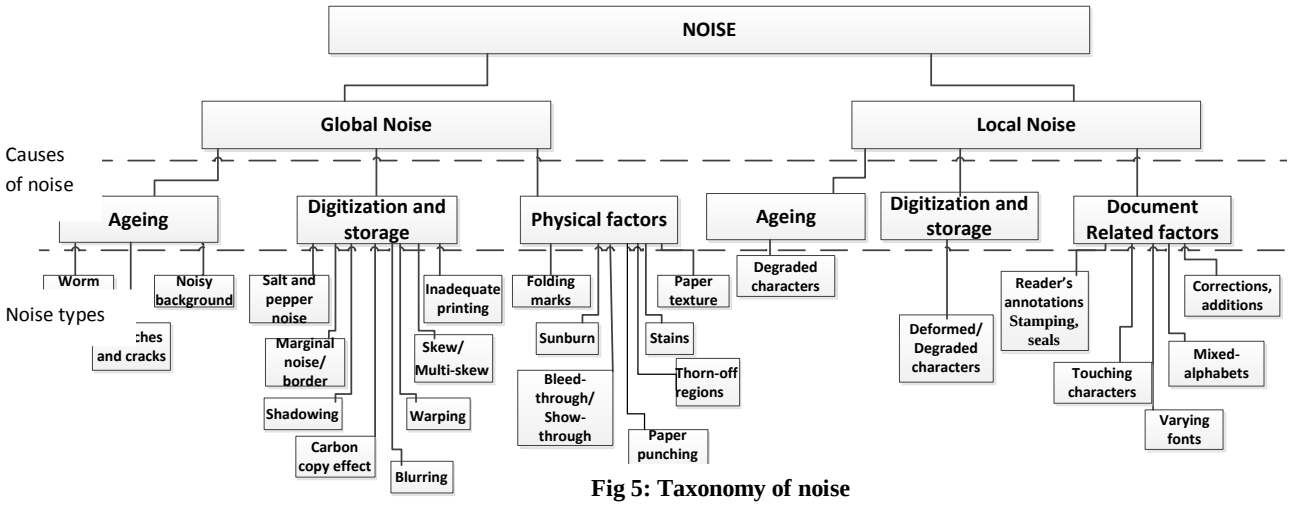

Fig 5: Taxonomy of noise

The histogram of noise occurrence taking 20 samples in each dataset is shown in Fig 10. Fig 11 shows the cumulative histogram of all the noise types in 4 datasets. It is clear from the cumulative histogram that noisy background, border, skew, touching characters and degraded characters are the prominent noise types. Fig 12 shows the histogram of noise based on their source.

As the dataset1 includes the most aged documents it will have all the noise related to the ageing factors. The aged documents will also have noise introduced during digitization. In historical documents we usually do not find noise like inadequate printing, punching, thorn-off regions, shadowing, carbon copy effects and sunburn. Since most of them the punching machine, printers, staplers, Xerox machine, thermal papers etc. were not used in olden days.

Hence both Fig 10 and 11 do not have any occurrence of these noise types. The document related noise is found in all datasets since it depends on the language and representation of the document. We do not find the physical factors as the cause for any prominent noise type.

In Fig 12 we can visualize that the datasets 1 to 4 are in chronological order. In Fig 12, histograms for ageing and physical factors are decreasing with respect to the chronological order.

Table 2 gives the detection, removal and restoration techniques for the prominent noise types. The detection techniques are methods which find where exactly in the image the noise is found. The removal techniques are methods which remove the noise. The restoration techniques are methods which are used to restore the document content in case the removal technique affects them.

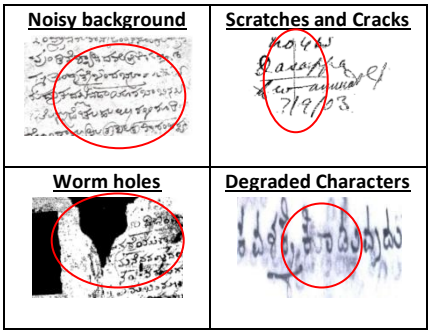

Fig 6: Types of ageing noises

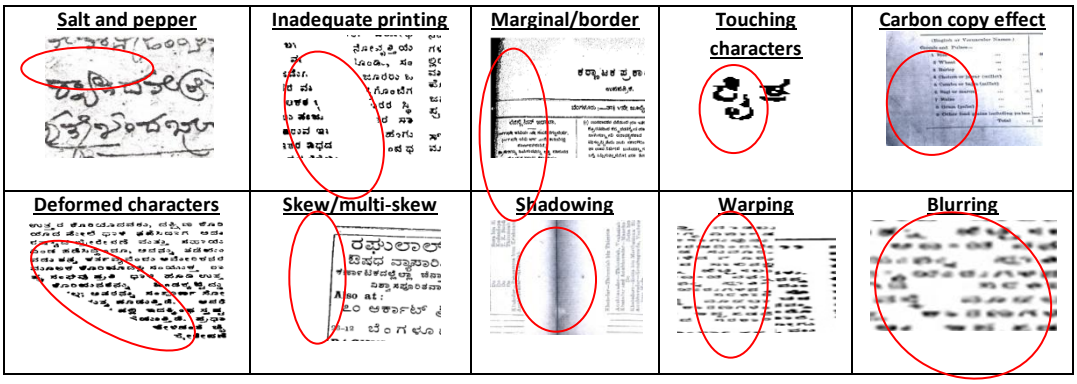

Fig 7: Types of digitization and storage noise

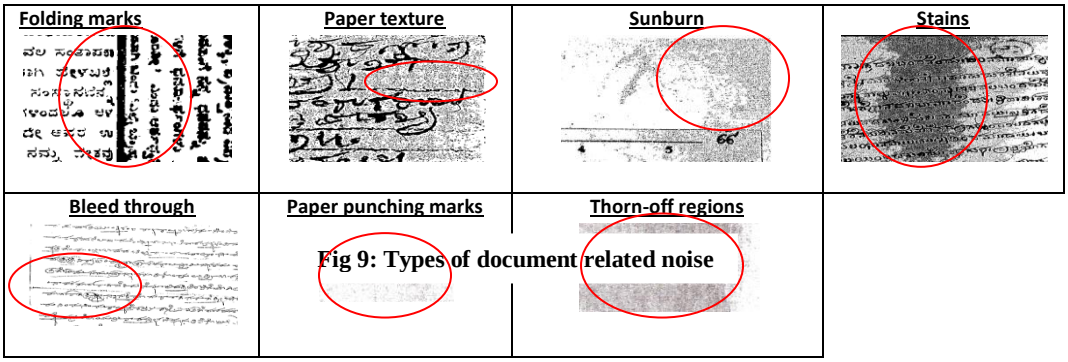

Fig 8: Types physical noise

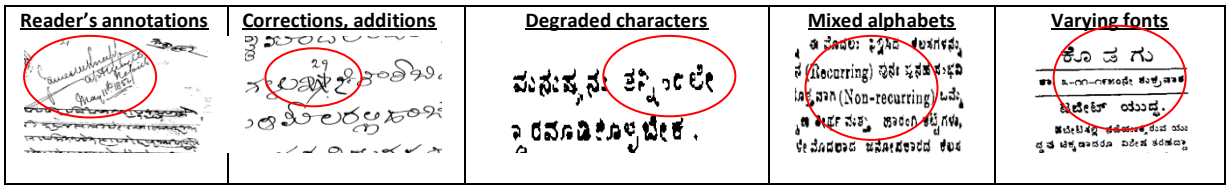

Fig 9: Types of document related noise

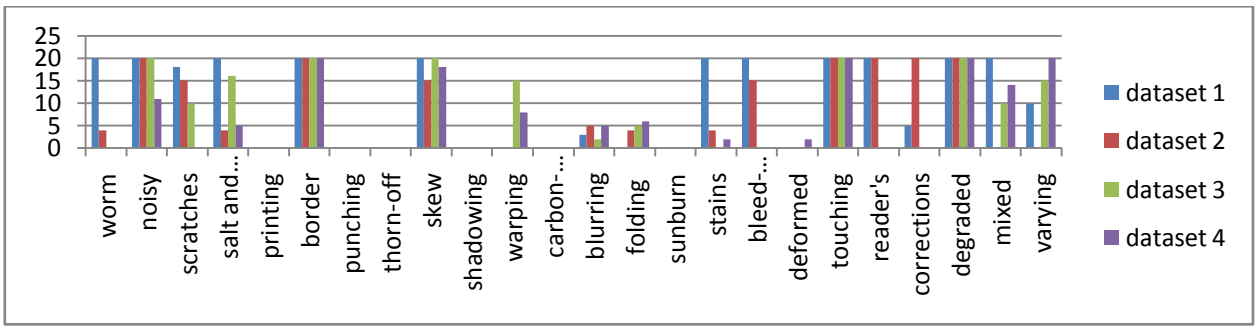

Fig 10: Histogram of noise in chosen datasets

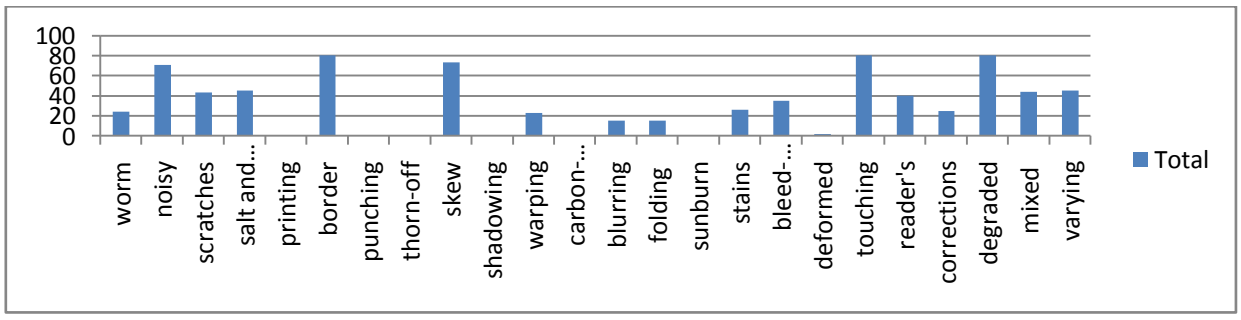

Fig 11: Histogram of occurrence of noise in the entire collection of 80 samples [20 samples in each of the four Datasets]

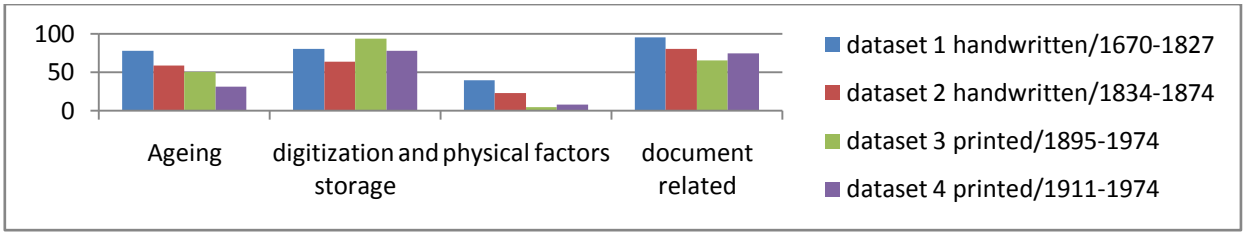

Fig 12: Histogram of noise based on the source for each dataset

3.1 Border

The border noise is the global noise due to the page surroundings which has noise, sometimes the overlap of text of the next page, the folding marks of the book, photocopy/scan effect etc. All these factors are because of digitization. The noise will be introduced either in border, center or corner of the page. Fig 13a shows a sample of border noise in surroundings and 13b shows a sample noise in the center of the document.

Fig 13a: Border noise in the surroundings of English

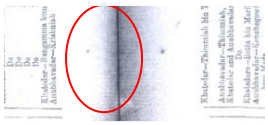

Fig 13b: Border noise in the center of English document

Techniques for border noise detection and removal

The border noise is mostly due to digitization process. Some of the detection and removal techniques are as follows.

- Zheng Zhang et.al[10] proposes a method to remove the border noise and correct the warped words is proposed. It uses the projection profile to detect the border noise and modified Niblack's method to remove the shadow. Connected component analysis is used to restore and adjust the location and orientation of the warped word in the shadow area.
- Worapoj Peerawit et.al[11] Proposes use of Sobel detector and the edge density property to detect the edges to differentiate noise and text areas and removes the edge noise.
- Syed Saqib Bukhari et.al[12] Presents a method to remove the border noise in a camera captured image. The page frame detection technique which uses text and non-text contents information to find the page frame of document images.

3.2 Skew

Skew is a prominent type of noise introduced in the digitization process. The main factor causing this noise is the paper being not placed properly during scanning/photocopying. Generally handwritten documents written on an un-ruled page will have skew. Skew is a global noise since it manifests itself throughout the page. Fig 14 is a sample for skew.

Techniques for skew detection and correction

There is a lot of work reported in the literature addressing skew detection and correction. Some of the techniques reported are:

- B.V.Dhandra et.al[1] Proposes image dilation and region labeling technique for estimation of skew angle in a binary document image. The labeling of the region is done using depth first search. The orientation angle is calculated for each labeled regions. Then the skew angle is calculated considering the average of all orientation angles. However the method works only for machine printed documents.
- Manjunath Aradhya.V.N et.al[2] Proposes method to estimate the skew angle. Some characters are blocked and thinning is done on these blocked regions. These thinned coordinates are fed to the Hough transform (HT) to estimate the skew angle.
- D.R.Ramesh Babu et.al[4] Proposes estimation of skew angle based on finding the centroid of each connected component (CC) and plotting the major axis of the ellipse of each connected component.
- Avanindra et.al[8] is based on interline cross-correlation in the scanned image. It is calculated over small regions selected randomly. The maximum median of cross-correlation is used as the criterion to obtain the skew, and a Monte Carlo sampling technique is chosen to determine the number of regions over which the correlations have to be calculated.
- S. Banerjee et.al[9] Proposes a page edge detection algorithm that is a multiplicative combination of gradients and line based page edge detectors, a skew detection algorithm that is a linear combination of page/content edge and content based predictors, and a pipeline for skew correction and frame removal.
- H. Fan et.al[13] Proposed a Rectangular Active Contour Model (RAC Model) for content region detection and skew angle calculation by imposing a rectangular shape constraint on the zero-level set in the Chan–Vese Model (C-V Model) according to the rectangular feature of content regions in document images.
- Xiaoyi Jiang et.al[14] Describes an algorithm to estimate the skew angle of document images. It uses the nearest-neighbor clustering paradigm.

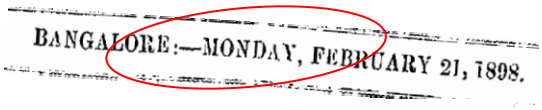

Fig 14: Sample for skew in English document

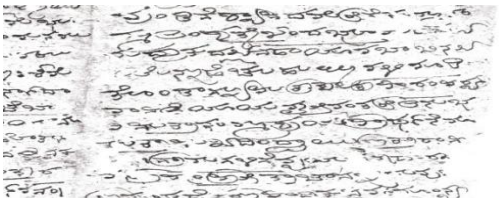

Fig 15: Sample for noisy background in Kannada document

3.3 Noisy background

Noisy background is a major noise type in all historical documents. All the historical paper documents will be aged and this changes the color of the paper to yellowish. The color causes noisy back ground in digitized usage. Fig 15 is a sample for noisy background.

Techniques for noisy background detection and removal

Since noisy background is present throughout the document the detection and removal techniques have to also identify the content.

- Negishi.H et.al[15] Proposes a method for separating characters from noisy background using background equalization and automatic thresholding. The word extraction is done using connected components and labeling.
- Abhijit Mitra et.al[16] Proposes a new thresholding method by maximizing the ratio of standard deviations of the combined effect on the image to the sum of weighted classes and finally the image restoration phase by image binarization utilizing the proposed optimum threshold level.
- Wafa Boussellaa et.al[17] Proposes a method in which document image is considered as a mixture of Gaussian densities representing the foreground and background document image components. The EM algorithm is used to estimate and improve the parameters of the mixtures of densities recursively. The initial parameters of the EM algorithm are estimated by the k-means clustering method, and then the document image is partitioned into text and background classes by the means of Maximum Likelihood approach.

3.4 Touching Characters

The Kannada language has touching characters as a pattern. They are called as conjuncts or vowel modifiers. These characters are difficult to be segmented as they neither have separation vertically nor horizontally. Fig 16 shows an example for the touching characters.

Techniques for segmenting touching characters

In literature less work is reported on segmenting touching characters.

- B.M. Sagar et.al[3] Proposes a method for segmentation using brute force method. The character segmentation and line segmentation is done using the projections. The conjuncts are separated using connected components approach.
- Bansal, V[6] Proposes an algorithm for the structural properties of the script. The width and height of each character is calculated and used for identifying the touching characters.

3.5 Degraded characters

The historical documents undergo severe degradation like ink fading, deterioration of paper etc. with time. Due to this the strokes of the characters are lost. Some other causes are due to the digitization process like bleaching, binarization etc. where the thresholding value of the noise and the character will be almost equal removes even the strokes of the characters. Fig 17 is a sample for the bleached historical document having degraded characters.

Techniques for degraded character detection and restoration

Deformation is sometimes due to the warping and skew. The restoration methods have to remove skew first in such cases.

- Minoru Mori et.al[18] is a run-length compensation technique is used for extracting approximate directional run-lengths of strokes from degraded handwritten characters and this technique is applied to the conventional feature vector based on directional run lengths to identify the degraded characters.
- Reza Farrahi Moghaddam et.al[7] Proposes a method which firstly uses the multi-level classifiers, which take prior information of the stroke width which allows locating candidate character pixels. Secondly, a

  level set active contour scheme is used to identify the boundary of a character which finds the degraded characters.

Fig 16: Sample for touching characters in kannada document

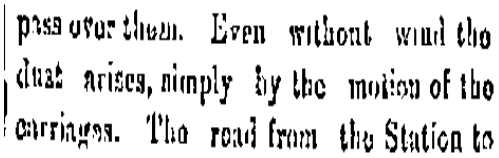

Fig 17: Sample for degraded characters in English document

5. CONCLUSION

A taxonomy of noise for degraded scripts is proposed. This taxonomy is language independent and will hold good for historic documents in any language. Five prominent noise types are identified based on the collected dataset. We have catalogued the related techniques for detection, removal and restoration for each of these prominent noise types namely: border, skew, noisy background, touching characters and degraded characters. There is scope for research on the noise types which have not been addressed in the literature such as the scratches and cracks, worm holes, inadequate printing, paper texture, sunburn, stains, punching marks, thorn-off regions, reader's annotations, corrections and additions, mixed alphabets, varying fonts identified in possible noise types. Our vision is to build an OCR system for digitizing and preserving historical Kannada documents. In this context, we

intend to further research and develop methods to detect, remove and restore the prominent noise types like touching characters and degraded characters for which not much is reported in literature.

Table 2. Prominent noise types and their handling techniques

Type of Noise	Source	Detection Technique	Removal Technique	Restoration Technique
Border/ Marginal noise	Digitization and Storage	Projection profile[10], Sobel Edge detection using edge density[11], Text line detection and Page frame detection[12]	Niblack's algorithm and Binarization[10], Removing black pixels after edge detected[11]	Connected component clustering and word based clustering[10] [restores the characters from getting degraded]
Skew	Digitization and Storage	Image dilation and region labeling[1], Hough transform[2], Connected components[4], Inter slice cross-correlation[8], Content based predictors[9], Rectangular active contour[13], Nearest neighbor clustering[14]	Rotation [9]	NA
Noisy background	Ageing	NA	Background equalization [15], Bi-level Adaptive thresholding and binarization[16], E-M based algorithm & M L segmentation [17]	Connected components on down scaled image[15]
Touching Characters	Document related noise	Structural properties[6]	Vertical projection, horizontal projection, divide the character into 3 parts[3]	NA
Degraded characters	Digitization and Storage Ageing	Run length compensation[18], Multi-level classifiers[7]	NA	Feature vector based on compensated run lengths[18], Active contour scheme[7]

6. ACKNOWLEDGEMENT

We would thank Karnataka state archives department for providing the datasets for historical document images.

7. REFERENCES

[1] B.V.Dhandra, V.S.Malemath, Mallikarjun.H, Ravindra Hegadi, Skew Detection in Binary Image Documents Based on Image Dilation and Region labeling Approach, The 18th International Conference on Pattern Recognition (ICPR'06), 2006.

[2] Manjunath Aradhya.V.N, Hemantha Kumar.G, Shivakumara.P, Skew detection technique for binary document images based on Hough transform, International Journal of Information Technology, Vol. 3,2006.

[3] B.M. Sagar, G. Shobha, P. Ramakanth Kumar, Character Segmentation Algorithms For Kannada Optical Character Recognition, Proceedings of the 2008 International Conference on Wavelet Analysis and Pattern Recognition, Hong Kong, 30-31 Aug. 2008

[4] D.R.Ramesh Babu, Piyush.M.Kumat, Mahesh.D.Dhannawat, Skew Angle Estimation and Correction of Hand Written, Textual and Large areas of Non-Textual Document Images: A Novel Approach, IPCV 2006, 510-515.

[5] R.D. Lins, A Taxonomy for Noise Detection in Images of Paper Documents - The Physical Noises. International Conference on Image Analysis and Recognition, LNCS 5627, pp 844-854. Springer Verlag, 2009.

[6] Bansal.V., and Sinha, R. M. K.Segmentation of touching and fused Devanagari characters. Pattern Recognition-2002. 35, 4, 875-893.

[7] Reza Farrahi Moghaddam, David Rivest-Henault, and Mohamed Cheriet, Restoration and segmentation of highly degraded characters using a shape-independent level set approach and multi-level classifiers, 10th International Conference on Document Analysis and Recognition, 2009.

[8] Avanindra, Subhasis Chaudhuri, Robust Detection of Skew in Document Images, IEEE transactions on image processing, vol. 6, no. 2, february 1997.

[9] S. Banerjee, S. Noushath, P. Parikh, S. Ramachandrula, A. Kuchibhotla, A. Sharma, Real-time embedded skew detection and frame removal, 17th IEEE International Conference on Image Processing (ICIP), 2010.

[10] Zheng Zhang, Chew Lim Tan**,** Recovery of Distorted Document Images from Bound Volumes, Proceedings of Sixth International Conference on Document Analysis and Recognition, 2001.

[11] Worapoj Peerawit and Asanee Kawtrakul, Marginal Noise Removal from Document Images Using Edge Density, ICEP2004, Phuket, Thailand, 2004.

[12] Syed Saqib Bukhari, Faisal Shafaity, Thomas M. Breuel, Border Noise Removal of Camera-Captured Document Images using Page Frame Detection, 4th International Workshop on Camera-Based Document Analysis and Recognition, Lecture Notes in Computer Science, Beijing, China, Springer, 9/2011.

[13] H. Fan, L. Zhu, and Y. Tang, Skew detection in document images based on rectangular active contour, International Journal on Document Analysis and Recognition, vol. 13, no. 4, pp. 261–269, 2010.

[14] Xiaoyi Jiang, Bunke, H, Widmer-Kljajo, D, Skew detection of document images by focused nearest-neighbor clustering, Proceedings of the Fifth International Conference on Document Analysis and Recognition,1999.

[15] Negishi.H, Kato.J, Hase.H, Watanabe.T, Character Extraction from Noisy Background for an Automatic Reference System, Proceedings of the Fifth International Conference on Document Analysis and Recognition, 1999.

[16] Abhijit Mitra, Restoration of Noisy Document Images with an Efficient Bi-Level Adaptive Thresholding, World Academy of Science, Engineering and Technology 18 2006.

[17] Wafa Boussellaa, Aymen Bougacha, Abderrazak Zahour, Haikal EL Abed, Adel Alimi, Enhanced Text Extraction from Arabic Degraded Document Images using EM Algorithm, 10th International Conference on Document Analysis and Recognition, 2009.

[18] Minoru Mori, Minako Sawaki, Norihiro Hagita, Hiroshi Murase, and Naoki Mukawa, Robust feature extraction based on run-length compensation for degraded handwritten character recognition, Proceedings of Sixth International Conference on Document Analysis and Recognition, 2001.

[19] Website of Kaleido software & services, online available: http:// kannadaocr.com/downloads/ KanScanUser Guide_v10b. pdf.