Ancient Indian Scripts Image Pre-Processing and Dimensionality Reduction for Feature Extraction and Classification: A Survey

Abhishek Tomar[#1], Minu Choudhary[*2], Amit Yerpude[*3]

[#] *M. Tech. Scholar, Dept. of Computer Science and Engineering*
Rungta College of Engineering and Technology
Bhilai (C. G.), INDIA

[*] *Reader, Dept. of Computer Science and Engineering*
Rungta College of Engineering and Technology
Bhilai (C. G.), INDIA

***Abstract*— Engravings, inscriptions and epigraphs were used from archaic era to preserve knowledge and the great sayings. Diversified scripts are used in writing languages across the globe, in such an environment it is necessary to understand the script and the languages in the images or document prior to selecting an apt character detection and document analysis or prehistoric inscription analysis algorithm. Numerous methods for the automatic character identification and script recognition have been recommended so far. This manuscript is a terse survey on the image prior-processing techniques, segmentation techniques and feature extraction and classification via dimensionality reduction techniques. Broad research has previously been done in this domain but the ancient inscription character recognition is still challenging and needs more efficient techniques. This review will serve as basis for the preliminarily to image pre-processing and the efficacy of dimensionality reduction approaches in feature and classification.**

***Keywords*—Optical Character recognition (OCR), Image Processing, Feature Extraction and Classification, Principal Component Analysis (PCA), Independent Component Analysis (ICA), Simultaneous Blind Source Extraction (SBSE).**

I. Introduction

The analysis of inscriptions on stones, rock pillars, temple walls, copper plates and other writing material is called Epigraphy, the most fascinating and informative studies. It relates to art of writing which provides us a method for preservation and communication of ancient traditions. Inscriptions serve as the foundation for remodeling the past and ancient civilizations. These epigraphs serve as foundation for reconstruction of social, cultural, archaeological, and historical relic. Mostly inscriptions are records of donation to temples. From the study of inscriptions the exhaustive genealogies and documented spiritual practices, political organization, and legal codes has been recovered [1] [2]. To date, at least 90000 stone epigraphs already discovered from different parts of India. The stone surface is often distorted by various kinds of noises such as cracks, scratches, voids, etc. And also, the same letter takes different shapes depending on the skill of inscriber. In order to have a scientific method, the first step is to produce alphabet fonts of ancient scripts. By studying the lexis, syntax, and forms of the inscriptions, linguists can build understanding how languages developed and where they were used.

II. Writing Systems and Scripts of The Globe

There are six famous writing systems. Logographic System, a symbol graphically represents a complete word. Syllabic System, every symbol represents a syllable, as used in Japanese. Alphabetic System, characters represents phonemes of a verbal language. Abjads is alike alphabetic system, but has symbols for consonantal noise only. Unlike most other scripts in the world, Abjads are written from right to left within a textline, this uniqueness is particularly helpful for identifying Abjad based scripts in pen computing. Abugidas is another alphabetic kind of writing system used by the Brahmic family of scripts that originated from the ancient Indian Brahmi script and includes almost all scripts of India and Southeast Asia. The last significant writing system is the featural system in which the symbols or characters represent the features that frame the phonemes [3].

III. Scripts Of India

The Brahmi script is vital writing systems due to its time depth and influence. It represents the initial post-Indus texts, and some of the inital historical inscriptions found in India. It is the predecessor to most of the scripts originated in South, Southeast, and East Asia. In Brahmi each indication can be either a clear-cut consonant or a syllable with the consonant and the inherent vowel /a/. Brahmi and Brahmi-derived scripts indicates the same consonant with a dissimilar vowel by sketching spare strokes, called *matras*, attached to the character. Nexus are used to indicate consonant clusters [4].

IV. Complexity of Indian Scripts Character Recognition

Issues in character recognition can be classified in to varied categories such as broken characters, overlapped characters, characters touching each other, skewed characters, uneven character intensity, alphabet size, two-dimensional structure,

inter-class similarity, symbol order variations, stroke order variations spanning multiple symbols, stroke order, number and direction variations within symbols [5],prime complications are discussed below.

A. Alphabet size

Most Indic scripts use huge number of characters whereas English uses less than 100. The internal graphic structure makes a split and conquers approach to recognition feasible in theory. However many consonant conjuncts are represented by visually divergent conjuncts bearing no semblance to the constituent consonant shapes (Fig. 1.). Combinations of consonant and vowel emerge new symbols which cannot be segregated into the base consonant and *matra*. Cursive styles and stroke and symbol order variations across different writers includes another challenges for this strategy.

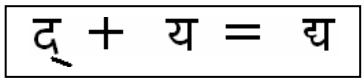

Fig. 1. Consonants forming distinct conjunct character in Devanagari

B. Two-dimensional structure

It is known *matras* or vowel can arise around the character or several components can surround the base consonant. Few likely vowel *matras* for a consonant symbol in Devanagari are shown in Fig. 2(a). In Fig. 2(b) & Fig. 2(c), shows a two-part *matra* with components in Tamil, half- consonant forms in consonant conjuncts around the base consonant and the vertical heap in case of Telugu respectively. Thus Indic scripts demonstrate a two-dimensional structure much, results in added unevenness in symbol and stroke order, not like the linear of the Latin script.

(a) (b) (c)

Fig. 2. Two-dimensional structure: (a) few likely *matras* for a consonant in Devanagari (b) two-part *matra* surrounding the consonant in Tamil (c) consonant conjunct in Telugu

C. Inter-class similarity

In Indian scripts, there is inherently high inter-class likeness among some pairs of symbols. Fig. 3(a) shows two characters from Malayalam, the only difference is small loop present in the first. Fig. 3(b) shows two Tamil characters with a slight difference in *matras*. This needs extremely distinctive features to describe the shapes of characters and graphemes.

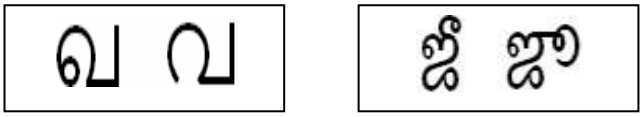

Fig. 3. Similar-looking characters in (a) Malayalam (b) Tamil

D. Issues with writing styles

While writing a character, users are usually anxious with reconstructing its visual appearance rather than its phonological structure. A variety of factors such as the relative positions of diverse strokes, the effort required to move from one stroke to the next given the overall flow of writing, and writing styles taught in school all have an influence on the stroke order that is eventually used. The consequences for Character Recognitions are many:

1) Symbol order variations: The sequence of writing of consonant and vowel units in a character need not correspond to the phonological order of their occurrence in the corresponding syllable. For instance, '◌ि' matra in Devanagari and '◌ெ' in Tamil are often written before writing the base consonant, since they occur to its left. In contrast, the Unicode representation of characters in Indic scripts is based on their phonological structure, and encodes the consonant before the vowel. While modelling characters and the lexicon, the recognition system should take this inconsistency into account.

2) Stroke order variations spanning multiple symbols: Strokes from different graphemes may be interleaved while writing a character. For example, a two-stroke matra may be written partially and completed only after the base consonant is written. This is slackly related to the phenomenon of delayed strokes in English, wherein a few strokes are entered only after the completion of the whole word. However for Indian scripts it happens at the character level and the variations are widespread and not limited to a small number of strokes such as t-crossings and i-dots in English.

3) Stroke order, number and direction variations within symbols: The ordering of strokes is likely to vary even within a symbol. In general, stroke order, number and direction variations are quite high in Indic characters and constitute one of the central challenges in online recognition of Indic scripts. These variations may be discovered involuntarily from the data samples by applying unsupervised learning techniques such as clustering.

E. Language-specific and regional differences in usage

Noticeable differences in the use of symbols may be observed in the use of a script like Devanagari across languages such as Hindi, Sanskrit, Marathi and Nepali. For instance, the *halant* (vowel muting diacritic) is used often in Sanskrit, but rarely in Hindi. The shape of symbols shows regional variations, prejudiced by other languages and scripts in use in the region and its nearby areas. In all cases, the language models, which are often used in the recognition system to improve accuracy, need to be considerably unique. The challenges for recognition of Indic scripts are adequately different from those for Latin. For instance, small vowel modifiers may get interpreted as noise in the input, and the *shirorekha* which is often written after completing the word requires special treatment (Fig. 4.).

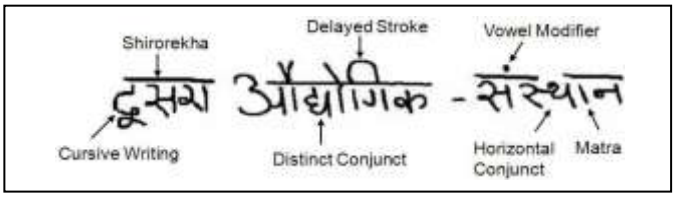

Fig. 4. Few challenges for Character Recognition of Devanagari script

V. Basic Steps For Character Recognition System

Character Recognition deals with the problem of identifying optically processed characters. Optical Character Recognition (OCR) can be performed either offline or online. The more constrained the input is, better the performance of the OCR system will be. However, unconstrained handwriting is still a challenge. A usual OCR system comprise of several components. In Fig. 5 a general setup is illustrated [6].

A. Optical scanning

Via scanning digitized image of original document is obtained. Optical scanners convert illumination intensity into gray-levels. Printed documents usually consist of black print on a white background. Hence, when performing OCR, it is common practice to convert the colour image into a binary image. Often this process, known as thresholding, is performed on the scanner to save memory space and computational effort.

B. Segmentation

Image Segregation or Segmentation is the procedure of segregating a digitized image into set of sub images. Segmentation outcome is much substantial and straightforward to analyze, and is normally used to locate objects and margins (lines, curves, etc.) in images. It is the procedure of conveying a tag to every pixel in an image such that pixels with the same tag possess assured characteristics and results a set of segments that communally form the intact image, or a set of contours extracted from the image. The prime issues in segmentation are:

- Touched and fragmented characters due to which it can be interpreted as one single character or that a piece of a character is believed to be an entire symbol.
- Dots and accents may be mistaken for noise, and vice versa.
- Inaccuracy in graphics or geometry for text or vice versa.

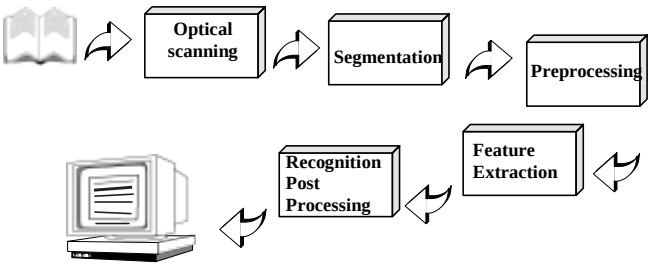

Fig. 5. Components of an OCR-system

C. Pre-processing

Scanned images may contain a definite amount of noise per as the resolution of the scanner and the success of the applied technique for thresholding, the characters may be messy or broken. Defects may cause deprived recognition rates, can be removed by using a pre-processor to smooth the digitized characters. The smoothing implies both filling and thinning. Filling eliminates small cracks and holes in the characters, while thinning reduces the thickness of the line.

D. Feature extraction

This phase incarcerates the essential features of the symbols. The techniques for extraction of such features are often divided into three main groups, where the features are found from, the distribution of points, transformations and series expansions, structural analysis.

E. Classification

The classification is the process of identifying each character and assigning to it the correct character class. First method is decision based recognition. These techniques are used when the portrayal of the character can be numerically represented in a feature vector. The principal approaches to decision based recognition are minimal remoteness classifiers, statistical classifiers and neural networks. Characters can also be derived from the physical structure of the character which is not as easily quantified in such cases the relationship between the characteristics may be of importance while deciding class membership.

VI. Available Methodologies

This illustrates the various effective techniques developed in the field of the field of OCR. Now we will portray various image pre-processing techniques to perform denoising, or filtering, binarization and segmentation techniques. Later we will depict feature extraction and classification techniques for extracting various features of the characters and then various classification techniques to recognize the characters.

There are many challenges addressed in handwritten document, ancient manuscripts and epigraphs or inscriptions image binarization, such as faint characters, correlation of foreground and background, bleed-through and large background ink stains [7]. Denoising eliminate the noise, enhances the text characters quality and make the background uniform whereas binarization or thresholding converts the grayscale image to binary i.e. black & white

There are a variety of filtering techniques available to denoise the images; primarily we classify filters in to two types Spatial Filter Domain (Mean Filter, Median Filter, Weiner Filter, Conventional Adaptive Median Filter, Decision Based Median Filtering, Bilateral Filter) and Frequency Domain (Butterworth Low Pass Filter, Gaussian Low Pass Filter) ([8], [9], [10], [11]).

A. Filtering Mechanisms

1) Spatial Domain Filters: Linear filtering in spatial domain apply a filter with a summation of weights of adjoining pixels. The weight is defined by the filter. Filtering is attained by convolution and convolution kernel is the correlation kernel rotated by 180^0.

a) Mean Filter: Mean filter is a linear filter. The intensity of every pixel in the image is replaced with the mean value of intensity of its adjoining pixels. The new value of intensity of a pixel (i, j) of an image I is given by:

$$I(i,j) = \frac{1}{M} \sum_{(x,y)\in N} I(x,y) \quad (1)$$

Where M represents the number of neighborhood pixels in N.

b) Median Filter: Median filter is a non linear filter. The median filtering technique, uses complex technique to detect the impulse noises and then filter the pixels. For $A\{a_1, a_2, a_3, \ldots . a_n\}$ and $a_1 \leq a_2 \leq a_3 \leq \cdots \leq a_n \in R$ the new value of intensity of a pixel (i, j) of an image I is given by:

$$\text{median}(A) = \begin{cases} a_{\frac{n+1}{2}}, \text{if n is odd} \\ \frac{1}{2}\left(a_{\frac{n}{2}} + a_{\frac{n}{\frac{n}{2}+1}}\right), \text{if n is even} \end{cases} \quad (2)$$

c) Bilateral Filter: Bilateral filter is a non linear filter in spatial domain, which does averaging without smoothing the edges. The bilateral filter takes a summation of weights of the pixels in a local neighbourhood; the weights depend on both the spatial distance and the intensity distance. Actually the bilateral filter has weights i.e. a product of two Gaussian filter weights, one of which corresponds to average intensity in a spatial domain, and second weight corresponds to the intensity difference. Hence no smoothing occurs, when one of the weights is close to 0. It means, the product becomes negligible around the region, where intensity changes rapidly, which represents usually the sharp edges. As a result, the bilateral filter preserves sharp edges. Mathematically, at a pixel location x, the output of a bilateral filter is calculated as follows:

$$\tilde{I} = \frac{1}{C}\sum_{y\in N(x)} e^{\frac{-\|y-x\|^2}{2\sigma_d^2}} e^{\frac{-|I(y)-I(x)|^2}{2\sigma_\gamma^2}} I(y) \quad (3)$$

Where σ_d and σ_γ are parameters controlling the drop of weights in spatial and intensity domains, respectively. I and $\tilde{I}$ are input and output images respectively. $N(x)$ is a spatial neighbourhood of pixel $I(x)$ and C is the normalization constant:

$$C = \sum_{y\in N(x)} e^{\frac{-\|y-x\|^2}{2\sigma_d^2}} e^{\frac{-|I(y)-I(x)|^2}{2\sigma_\gamma^2}} \quad (4)$$

2) Frequency Domain Filters: Spatial frequency filtering is implemented by low pass filters which perform Fourier transform. These are smoothing frequency filters, since they smooth edges and sharp transitions in an icon, such as noise. Low frequencies in the Fourier transform of an image are liable for the grey level manifestation over the smoothed areas. On the other hand, high frequencies are responsible for the presence of details, edges and noise in the image.

a) Butterworth Low Pass Filter: Butterworth Filter is a low pass filter with transfer function:

$$H(u,v) = \frac{1}{1+[D(u,v)/D_0]} \quad (5)$$

Where D_0 is a specific non negative quantity and D (u, v) is the distance from point (u, v) to the centre of the frequency rectangle.

b) Gaussian Low Pass Filter: Gaussian low pass filter removes effectively the noise but blurs the image. The mathematical form for two- dimensional Gaussian filter is given by

$$H(u,v) = e^{-D^2(u,v)/2\sigma^2} \quad (6)$$

Where D (u, v) is the distance from origin of the Fourier transform.

B. Segmentation

Image Segmentation is broadly classified in to types Local Segmentation and Global Segmentation. Local Segmentation deals with the sub images which is a small part of an image whereas Global segmentation is concerned with segmenting an entire image ([12], [13]). Here we will discuss two broad categories of thresholding, global thresholding and local thresholding [14].

Global Thresholding is the simplest implementation of thresholding is to choose an intensity value as a threshold level and the values below this threshold become 0 (black) and the values above this threshold become 1 (white). If T is the global threshold of image f(x,y) and the g(x,y) is the thresholding image, then:

$$g(x,y) = \begin{cases} 1, if\ f(x,y) \geq T \\ 0,\ otherwise \end{cases} \quad (7)$$

Among the global techniques the most efficient is Otsu's technique. Otsu's method applies clustering analysis to the grayscale data of input image and models two clusters of Gaussian distribution of pixels of the image. The optimal threshold minimizes the class variance of the two classes of pixels. But the common problem in document images are changes in illumination, or local shadows that are difficult to give a global threshold, i.e. for the whole image in such cases locally adaptive thresholding mechanisms are efficient.

Locally Adaptive Thresholding Methods, where a threshold that is computed at each pixel belongs to this class of algorithms. The threshold value depends upon range, variance, and surface fitting parameters or their logical combinations. It is typical of locally adaptive methods to have several adjustable parameters. The threshold T (i, j) is a function of i, j; the object or background decisions at each pixel is represented by the variable B (i, j) some of these methods, Yasuda, Niblack, Palumbo, White, Bernsen, Yanowitz, Kamel, Oh, Sauvola ([13], [15]) some are described below.

a) Niblack: This method [16]adapts the threshold as per to the local mean and standard deviation on the basis of a window size of b x b. The threshold at pixel (i,j) is calculated as:

$$T(i,j) = m(i,j) + k.\sigma.\sigma(i,j) \quad (8)$$

where m(i, j) and $\sigma(i,j)$ are the local sample mean and variance, respectively.

b) Bernsen: In this techniques the value of threshold is set to the average value of the minimum and maximum of a local window. Thus one has:

$$\begin{aligned} T(i,j) &= 0.5\,[\max_w(I(i+m,j+n)) + \min_w(I(i+m,j+n))] \\ &= 0.5\,[I_{high}(i,j) + I_{low}(i,j)] \end{aligned} \quad (9)$$

where w is a window of size b x b around the centre point (i, j). However if the contrast $C(i,j) = I_{high}(i,j) - I_{low}(i,j)$ is below a certain threshold (this contrast threshold was 15) then that neighbourhood is said to consist only of one class, print or background, depending upon the value of T (i, j). The window size is chosen as w = 31.

c) Sauvola: This technique claims to enhancement on the Niblack's technique particularly for blemished and inadequately illuminated documents. It adapts the threshold according to the local mean and standard deviation over a window size of bxb. The threshold at pixel (i, j) is calculated as:

$$T(i,j) = m(i,j) + [1 + k.(\frac{\sigma(i,j)}{R} - 1)] \quad (10)$$

where m(i, j) and $\sigma(i,j)$ are as in Niblack, and Sauvola proposed the values of k = 0.5 and R = 128. Thus the involvement of the standard deviation becomes adaptive. For case printed text on a dirty or stained paper the threshold is lowered. Sauvola's method accuracy of 94.9% followed by

Niblack's 93.7 % followed by Bernsen's 79.7 % [17], Sauvola's methods proves best for the spots and stains, shadows or wrinkles, Ink seeking from other side of document, Red colour characters [9].

Chew Lim Tan and Shijian Lu [18] proposed technique first estimate the shading divergence by two stages of polynomial surface smoothing process. First stage detects the blank background by fitting a polynomial surface PS_f to the intensity of pixels within the entire image. The second estimates the shading divergence by correcting a polynomial surface PS_s to the intensity of pixels that have been categories to the background. After the fitting of PS_f, pixels are alienated into two categories, an inked category that contains the pixels of text and other dark document components and a background grouping containing document pixels within the blank regions. The method reaches capable of 91.33%, which is much superior to Otsu's and Sauvola's (61.58% and 78.34%). Niblack's method can attain a 90.29% segmentation rate, but it generates a huge number of background noises which will hamper the consequent document processing tasks such as OCR.

Bolan Su et al. [19] presented technique first performs an adaptive contrast map for an input despoiled document image. The contrast map is then binarized and pooled with Canny's edge map to identify the text stroke edge pixels. The document text is then segregated by a local threshold that is anticipated on the basis of intensities detected for the text stroke edge pixels within a local window. The method achieves accuracies of 93.5%, 87.8% and 92.03%. The PSNR values Otsu's 15.34, Sauvola's 16.39, Niblack's 9.89, Bernsen's 8.89, Proposed method 19.65.

T Kasar et al. [20] presented a technique, employs an edge-based connected component approach and involuntarily determines a threshold for each component, has several merits over existing binarization methods. Firstly, it can deal with multiband texts with diverse background shades. Secondly, it can handle text of broadly anecdotal sizes, usually not handled by local binarization methods. Thirdly, the involuntarily computation of threshold for binarization and the logic for complementing the output from the image data and does not need any input parameter. This technique is found to have a fine compliance.

Document image binarization is a complicated task, particularly for complex document images. Non-uniform background, stains, and deviation in the intensity of the printed characters are some form of challenging document features. David Rivest-Hénault et al. [21] presented a local linear level set technique for the binarization of degraded historical document images. Here binarization is consummate by taking gain of local probabilistic models and of a flexible active contour scheme. More specifically, local linear models are used to guesstimate both the anticipated stroke and the background pixel intensities, and then used as the prime motivation in the proliferation of an active contour. In addition, a curvature-based force is used to control the viscosity of the contour and leads to more natural-looking results.

The main issue of extracting independent components (ICs) is to learn the de-mixing matrix from the known training images which can be outspread to vectors in conventional independent component analysis (ICA). However, the outspread vectors lead to the Small Sample Size problem (SSS) and the annoyance of dimensionality. Quanxue Gao et al. [22] proposed method solves these issues by encoding each input image as a matrix. In comparison with the conventional ICA, this method directly evaluates the two correlated de-mixing matrices from the image matrix without matrix-to-vector transformation, greatly alleviates the SSS and the annoyance of dimensionality, reduces the computational complexity, and concurrently utilizes the spatial and structural information embedded in image. Extensive experiments prove that this technique is better than standard ICA technique and some unsupervised methods.

Indu Sreedev [23] et al. showed that Natural gradient Flexible ICA (NGFICA) is a appropriate technique for segregating signals from a mixture of extremely correlated signals, as it reduces the reliance among the signals by considering the slope of the signal at each point. The proposed technique enhances word and character recognition accuracies of the OCR system by 65.3% (from 10.1% to 75.4%) and 54.3% (from 32.4% to 86.7%), respectively. Three ICs of the coloured image were attained by performing NGFICA on the extracted R, G, and B components. NGFICA output image can be considered as foreground, background, and noise images. Comparison of average threshold of each of these output images is done with that of original image. The one which is farther than original image is identified as the foreground as it had only text then the edge detection is done via Sobel edge detection and then dilation via disc-shaped structuring element to recover the characters.

Ayush Tomar et al. [24] proposed and effective image enhancement techniques for historic inscription images. ICA based techniques can become computationally very luxurious when the number of source signals is at large say 150 or more. Simultaneous Blind Source Extraction (SBSE) overcomes this issue by providing a provision and ability to extract the preferred number of independent components (ICs) from a collection of linear mixtures of huge amount of statistically independent cause signals. The approach is to use a contrast function to deal with the third and fourth order cumulants simultaneously to reduce the computational time overhead. This technique uses a contrast functional that captures higher order cumulants, which is maximized by the blind source extraction procedure to ICs. The final binarized image is obtained by computing a suitable local threshold level as per Otsu's thresholding and the post-processing via morphological operations dilation and erosion were to improve the readability of text in the text layer followed by a suitable median filter.

Saxena, Lalit Prakash [25] presented an effectual binarization technique for readability enhancement of degraded manuscripts. A stain is a superficial colour change as lying within the manuscript fibers. Eliminating stains leads to enhanced manuscript image quality and hence improved readability. The technique is tested on manuscript images, from distinct media having diverse scripts: palm leaf (Grantha), rock (Brahmi), and paper (Modi, Newari, Persian and Roman), and Document Image Binarization Contest datasets (DBICO). This technique attains 66.27%, 92.15%, 97.90%, 56.23%, 78.62% and 98.91% readability for rock, palm leaf and paper manuscripts respectively. For the Brahmi manuscript images this technique handled the low contrast and low illumination efficiently, in any case, for most of the image region. The

resultant has clean symbols with no background noise and good intra-connectivity, whereas the noises at some regions conceal the symbols. There is a ostensible presence of blurred corners of the thresholded image attained by the techniques and the corner symbols are thicker than the symbols present in the original image. This technique effectively eliminates all the stain effects and the intra-connectivity is maintained too.

C. Dimensionality Reduction for Feature Extraction and Classification Techniques

Real life data, such as speech signals, digital photographs, or fMRI scans, generally has a high dimensionality. To deal this data effectively, its dimensionality needs to be reduced. Dimensionality reduction is the alteration of high-dimensional data into a significant depiction of diminished dimensionality. Ideally, the diminished depiction should have a dimensionality that corresponds to the inherent dimensionality of the data. The inherent dimensionality of data is the minimum number of parameters needed to account for the observed properties of the data [26]. Dimensionality reduction is important in many domains, since it mitigates the pest of dimensionality and other adverse properties of high-dimensional spaces [27].

Conventionally, dimensionality reduction was performed using linear technique called Principal Components Analysis (PCA), this linear technique cannot effectively handle complex nonlinear data. Therefore, a huge number of nonlinear methods for dimensionality reduction have been suggested. In comparison to the conventional linear methods, the nonlinear methods have the capability to deal with complex nonlinear data. Especially for real-world data, these nonlinear dimensionality diminution techniques may propose an advantage, because real-world data is prone to be highly nonlinear.

The prime distinction between linear and nonlinear techniques is linear techniques presume that the data lie on or near a linear subspace of the high-dimensional space so they perform dimensionality diminution by implanting the data into a subspace of lower dimensionality. Although there are diverse techniques to do so, PCA yet the most popular (unsupervised) linear technique. Nonlinear methods for dimensionality does not depend on the linearity hypothesis as an outcome, more complex implanting of the data in the high-dimensional space can be seen [28].

1) Principal Component Analysis (PCA):

PCA [29]constructs a low-dimensional depiction of the data that describes as much of the variance in the data as possible. In mathematical terms, by calculating the eigenvectors of the covariance matrix of the original inputs, PCA linearly transforms a high-dimensional data into a lower dimensional space which are uncorrelated and orthogonal. PCA first compute the Eigen values of the feature matrix; sort the Eigen values and ignore the really small values then transform the data into the Eigen space formed by the selected Eigen vectors.

2) Kernel based Principle Components Analysis (KPCA):

KPCA [28] is a non linear PCA twisted using the kernel trick. KPCA maps the original inputs into a high dimensional feature space using a kernel method. Mathematically, current features are transformed into high-dimensional space and the calculate eigenvectors are transformed in this space. Discard the vectors with really low Eigen values and then perform learning in this transformed space. KPCA is computationally intensive and takes much time compared to PCA the reason being that the number of training data points in KPCA is much higher than PCA. So number of principle components that need to be estimated is also much larger.

3) Independent Components Analysis (ICA):

The common scenario of ICA is the blind source extraction problem where the goal is to recover mutually or communally independent but unknown source signals from their linear mixtures without calculating the mixing coefficients. Two distinctions between PCA and ICA are that the components are statistically autonomous and not uncorrelated and second, the un-mixing matrix is not orthogonal like PCA [30]. The algorithm works on the principle of minimizing communal information between the variables, minimizing communal information is the correct criteria for judging independence. Also minimizing communal information is same as maximizing entropy. There are several algorithms for doing ICA one of the most popular ones being FastICA [31].

Utpal Garain et al. [32] proposed an ICA based enhancement technique to improve the precision for machine reading. Images of inscriptions that are usually carved on stones or other robust materials and found at the sites of historical monuments are taken as an input for conducting the experiments. Momentous improvement in classification rate of an OCR system proves the potential of the proposed ICA-based method. It enhance word and character detection accuracies of the OCR system by 68.6% (from 11.2% to 79.8%) and 57.3% (from 34.8% to 92.1%), respectively.

Keerthi Prasad G et al. [33] presented a system which is viable for real time applications. The proposed system is implemented on handheld device via two distinct ways i.e. PCA and Dynamic Time Wrapping (DTW). The attained outcome for PCA approach shows potential than DTW. On an average, up to 88% recognition rate is achieved for the PCA approach and up to 64% for DTW approach, also the time elapsed for recognition of unknown character is about 0.8 sec for PCA approach, and about 55 sec for DTW approach, thus the PCA approach is suitable for real-time applications.

Abdelmalek Zidouri [34] presented a PCA based classifier. Bi-cubic interpolation was utilized to interpolate the data. In this approach, recognition is applied at two levels. First stage recognition is to recognize secluded characters while second stage recognition is applied to segregated characters. The scheme was tested on different fonts found very interesting results. First stage recognition using weighted similarity match based upon pre-stored feature values, produced recognition of 98%. The recognition at second stage for non isolated characters resulted in 90% recognition; where misclassification was primarily because of segmentation errors.

Manjunath Aradhya V N and Hemantha Kumar G [35] proposed a new scheme, PCA in combination with Neural Network for character recognition. PCA extracts the desired number of principal components of multidimensional data. Generalized Regression Neural Network (GRNN), where it has radial basis layer and a special linear layer is used for

subsequent classification purpose. The method achieves 99.78 % accuracy clear and degraded, 94.5 % for noisy data.

Deepu V. et al. [36] proposed a PCA-based classification for online detection of secluded handwritten characters. The set of points from the digitized image is smoothed and normalized and a feature vector then extracted. Standard PCA is used to diminish dimensionality of each class and the orthogonal distance to the class subspace used for classification. Pre-clustering and alteration of the distance measure are explored as ways of addressing specific problems with PCA. The projected alterations are found experientially to improve recognition accuracy, and the resulting schemes found to compare auspiciously with the conventional Nearest Neighbour classifier. A remarkable aspect of these methods is that adding or replacing a training pattern of a specific class requires re-computation of the principal components of only that class. Another salient point is that no language or script specific features are used in the pre-processing, feature extraction or classification makes these methods extensively applicable for the recognition of other scripts.

Purnima Kumari Sharma et al. [37] projected a Radon Transform in combination with PCA, technique can be a very fast and efficient for Character Recognition. Via Radon Transform the characters can be rotated at any angle. Radon transformation is noise resistant which gives us a better image. PCA expresses the large 1-Dimensional vector of pixels constructed from 2-Dimensional character image into the dense principal components of the feature space. It gives the best fit vector for each character on which it is applied. The prime advantages of this technique are: finding geometric relations of the character by Radon Transform, invariance to background noise and low computational complexity. The method obtained around 85-95% accuracy experimented on 200 samples from different writers.

Delac Kresimir et al. ([38], [39]) did an independent, proportional study of three renowned appearance based face recognition algorithms PCA, ICA and Linear Discriminant Analysis (LDA) shows that PCA outperformed all others with illumination changes task.

Zhang Daohui et al. [40] did an experimental study, results showed that PCA had a deprived classification accuracy compared with the other two schemes, and the difference in classification accuracy of LDA and PCA+LDA was marginal. PCA is a good tool for dimensionality reduction, yet lacks of the ability of the class partitioning. LDA considers the class partitioning when it learns from the training samples to attain a linear optimal projected matrix. Thus, LDA had a higher accuracy than PCA. Though PCA+LDA did not bring complex computation to the projection process, it could not improve the ability of the pattern recognition much. Because the three projection schemes all obtained a projected linear matrix in the end, the processing time of them had no big difference.

Yunfei Jiang and Ping Guo [41] did a relative study of some feature extraction methods for face recognition in the identical conditions. The technique evaluated includes Eigen faces, KPCA, Fisher faces, Direct-LDA, Regularized LDA, and Kernel Direct Discriminant Analysis (KDDA). For comparison on feature extraction techniques, Nearest Neighbor (NN) algorithm is used, since this classifier is common and simple. Experiential studies was conducted to evaluate these feature extraction methods with images from ORL face database, and it has been observed that in most cases LDA-based methods are efficient than PCA-based ones. Although KPCA has an ability to describe the nonlinear problems, it does not always ensure better performance than Eigen faces. When the problem is linearly separable, performing KPCA on it might get even worse performance than Eigen faces. LDA-based methods surpass PCA-based ones on whole because of the inherent discrimination ability of the former, but it is not absolute. On ORL database, using the NN classifier for classification, KDDA outperforms the other five methods, and Fisher faces is the worst one.

VII. Conclusions

Among the filtering mechanisms median filtering is best suited for the salt and pepper noise but in some cases non linear bilateral filter proves best for enhancing the quality, it smoothes but with sharp edges. For segmentation sauvola's proves to be the best followed by niblack's for non-uniformly illuminated images. Sometimes these mechanisms also get failed like stained documents etc. then hybrid approaches like active contour, shading estimations, and forefront and background independent binarization etc. can produce better results. But these methods also have certain limitations in terms of inscriptions where background and foreground are highly simultaneous in such cases SBSE and NGFICA extracted independent components followed by morphological operations gives better result. LDA outperforms PCA and but in some circumstances like illumination changes PCA is best. ICA is best for the extraction for the blind source from the correlated signals whereas for handwriting based digit recognition the best results are obtained by KPCA, as compared to ICA which is followed by PCA. This demonstrates the fact that dimensionality reduction can improve the classification error and at the same time reduces the computation time of the learner. Though KPCA outperforms these two but has much principle components as compared to PCA and ICA. This is mainly due to the filtering of gratuitously attributes from the handwriting.

References

[1] Nalin Warnajith, Atsushi Minato and Satoru Ozawa Dammi Bandra, *"Creation of precise alphabets fonts of early Brahmi script from photographic data of Shri Lankan inscriptions,"* Canadian Journal on Artifical Intelligence, Machine Learning and Pattern Recognition, vol. Vol. 3, no. No. 3, pp. 33-39, 2012.

[2] Navya.K, Rajithkumar.B.K, Nagesh.C H.S. Mohana, *"Interactive Segmentation for Character Extraction in Stone Inscriptions,"* in 2nd International Conference on Current Trends in Engineering and Technology, ICCTET'14, IEEE Conference Number - 33344, Coimbatore, India., July 8, 2014, pp. 321-327.

[3] Tulika Dube, and Adamane P. Shivaprasad Debashis Ghosh, *"Script Recognition—A Review,"* IEEE transactions on pattern analysis and machine intelligence , vol. VOL. 32, no. NO. 12, pp. 2142-2161, DECEMBER 2010.

[4] brahmi.html. [Online]. http://www.ancientscripts.com/brahmi.html\

[5] Bharath A. and Sriganesh Madhvanath, *"OCR for Indic Scripts: Document Recognition and Retrieval,"* in Guide to OCR for Indic Scripts Advances in Pattern Recognition 2010. Hewlett-Packard Laboratories, Bangalore, India: Springer, 2010, pp. 209-234.

[6] Line Eikvil, *"Optical Character Recognition"*. P.B. 114 Blindern, N-0314Oslo: Norsk Regnesentral, December, 1993.

[7] B. Gatos, I. Pratikakis K. Ntirogiannis, *"A combined approach for the*

binarization of handwritten document images," Pattern Recognition Letters, vol. 35, pp. 3–15, 1 January 2014.

[8] C.R. and E.R. Woods Gonzalez, Digital Image Processing.: Prentice-Hall Inc., 2nd edition, 2002, pp. 75-278.

[9] Ventzas Dimitrios Ntogas Nikolaos, *"A Binarization Algorithm For Historical Manuscripts ,"* in 12th WSEAS International Conference on COMMUNICATIONS, Heraklion, Greece, July 23-25, 2008.

[10] Suman Shrestha, *"Image Denoising Using New Adaptive Based Median Filter,"* Signal & Image Processing : An International Journal (SIPIJ), vol. Vol.5 , no. No.4, pp. 1 -13, August 2014.

[11] K Srikanta Murthy, Arun Vikas Singh B Gangamma, *"Restoration of Degraded Historical Document Image,"* Journal of Emerging Trends in Computing and Information Sciences, vol. VOL. 3, no. NO. 5, pp. 792-798, May 2012.

[12] A. S. Chauhan and M. Dixit S. Silakari, *"Image Segmentation Methods: A Survey Approach,"* in Fourth International Conference on Communication Systems and Network Technologies, 2014.

[13] Mehmet Sezgin and Bu¨lent Sankur, *"Survey over image thresholding techniques and quantitative performance evaluation,"* Journal of Electronic Imaging , vol. 13, no. 1, pp. 146–165, January 2004.

[14] XIONG Fu-song, *"Survey over image thresholding techniques based on entropy,"* in International Conference on Information Science, Electronics and Electrical Engineering (ISEEE), 2014, pp. 1330-1334.

[15] Sudipta Roy, O.Imocha Singh, Tejmani Sinam and Kh.Manglem Singh T.Romen Singh, *"A New Local Adaptive Thresholding Technique in Binarization ,"* IJCSI International Journal of Computer Science Issues, vol. Vol. 8, no. 6, No 2, pp. 271-277, November 2011.

[16] W. Niblack, *An Introduction to Image Processing*, Prentice-Hall, 1986, pp. pp:115-116.

[17] M. PietikaK inen J. Sauvola, *"Adaptive document image binarization,"* Pattern Recognition, vol. 33 , pp. 225-236, 2000.

[18] Shijian Lu and Chew Lim Tan, *"Binarization of Badly Illuminated Document Images through Shading Estimation and Compensation,"* in Ninth International Conference on Document Analysis and Recognition, ICDAR 2007, vol. 1, 2007, pp. 312 - 316.

[19] Shijian Lu, Chew Lim Tan Bolan Su, *"A Robust Document Image Binarization Technique for Degraded Document Images,"* IEEE Transcation on Image Processing, vol. 22, no. 4, pp. 1057-7149, 2012.

[20] J Kumar and A G Ramakrishnan T Kasar, *"Font and Background Color Independent Text Binarization,"* in In Proceedings of 2nd International Workshop on Camera Based Document Analysis and Recognition, 2007, pp. 3-9.

[21] Reza Farrahi Moghaddam and Mohamed Cheriet David Rivest-Hénault, *"A local linear level set method for the binarization of degraded historical document images ,"* International Journal on Document Analysis and Recognition (IJDAR), vol. 15, no. 2, pp. 101-124 , 2012.

[22] Lei Zhang, David Zhang and Hui Xu Quanxue Gao, *"Independent components extraction from image matrix,"* Pattern Recognition Letters , vol. 31, pp. 171–178, 2010.

[23] Rishi Pandey, N. Jayanthi,Geetanjali Bhola and Santanu Chaudhury Indu Sreedevi, *"NGFICA Based Digitization of Historic Inscription Images,"* ISRN Signal Processing, vol. vol. 2013, p. 7 pages, 2013.

[24] Ayush Tomar, Aman Raj and Santanu Chaudhury S. Indu. (2014, November) http://vigir.missouri.edu/~gdesouza/Research/Conference_CDs/ACCV_2014/pages/workshop13/index.html. [Online]. http://vigir.missouri.edu/~gdesouza/Research/Conference_CDs/ACCV_2014/pages/workshop13/pdffiles/w13-p3.pdf

[25] Lalit Prakash Saxena, *"An effective binarization method for readability improvement of stain-affected (degraded) palm leaf and other types of manuscripts,"* CURRENT SCIENCE, vol. 107, no. 3, pp. 489-496, 10 AUGUST 2014.

[26] K. Fukunaga, *Introduction to Statistical Pattern Recognition*. San Diego, CA, USA,: Academic Press Professional, Inc., 1990.

[27] R. Kharal, *Semidefinite embedding for the dimensionality reduction of DNA microarray data*, 2006.

[28] EO Postma, HJ Van den Herik LJP Van der Maaten, *"Dimensionality reduction: A comparative review,"* Technical Report TiCC TR, 2009.

[29] N.P. Hughes and L. Tarassenko, *"Novel signal shape descriptors through wavelet transforms and dimensionality reduction,"* In Wavelet Applications in Signal and Image Processing , vol. X, pp. 763–773, 2003.

[30] K. Chua, W. Chong, H. Lee, and Q. Gu L. Cao, *"A comparison of pca,kpca and ica for dimensionality reduction in support vector machine,"* Neurocomputing, vol. 55, pp. 321-336, 2003.

[31] S. Marsland, *Machine Learning : An Algorithmic Perspective*, CRC Press, 2009.

[32] Utpal, Atishay Jain, Anjan Maity, and Bhabatosh Chanda Garain, *"Machine reading of camera-held low quality text images: an ICA-based image enhancement approach for improving OCR accuracy.,"* in In Pattern Recognition, 2008. ICPR 2008. 19th International Conference, 2008, pp. 1-4.

[33] G., Imran Khan, Naveen R. Chanukotimath, and Firoz Khan Keerthi Prasad, *"On-line handwritten character recognition system for Kannada using Principal Component Analysis Approach: For handheld devices,"* in In Information and Communication Technologies (WICT) 2012 World Congress on, 2012, pp. 675-678.

[34] Abdelmalek Zidouri, *"PCA-based Arabic Character feature extraction,"* in In Signal Processing and Its Applications, 2007. ISSPA 2007. 9th International Symposium on, IEEE., 2007, pp. 1-4..

[35] A. V. N., and K. G. Hemantha. Manjunath, *"Principal component analysis and generalized regression neural networks for efficient character recognition.,"* in In Emerging Trends in Engineering and Technology, 2008. ICETET'08. First International Conference on, IEEE, 2008., 2008, pp. 1170-1174..

[36] V., Sriganesh Madhvanath, and A. G. Ramakrishnan. Deepu, *"Principal component analysis for online handwritten character recognition,"* in In Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on, IEEE, 2004., vol. 2, 2004, pp. 327-330.

[37] Purnima Kumari, Mondira Deori, Balbindar Kaur, Chandralekha Dey, and K. Das. Sharma, *"Radon Transform and PCA based feature extraction to design an Assamese Character Recognition system,"* in Emerging Trends and Applications in Computer Science (NCETACS), 2012 3rd National Conference on, 2012, pp. 46 - 51.

[38] Delac Kresimir, Mislav Grgic, and Sonja Grgic. *"A comparative study of PCA, ICA, and LDA.,"* in In Proc. of the 5th EURASIP Conference focused on Speech and Image Processing, Multimedia Communications and Services, 205, pp. 99-106.

[39] Delac Kresimir, Mislav Grgic, and Sonja Grgic, *"Independent comparative study of PCA, ICA, and LDA on the FERET data set,"* in International Journal of Imaging Systems and Technology, vol. 15, 2005, pp. 252-260.

[40] Daohui, Xingang Zhao, Jianda Han, and Yiwen Zhao. Zhang, *"A comparative study on PCA and LDA based EMG pattern recognition for anthropomorphic robotic hand,"* in In Robotics and Automation (ICRA), 2014 IEEE International Conference on, IEEE, 2014, 2014, pp. 4850-4855.

[41] Yunfei, and Ping Guo. Jiang, *"Comparative studies of feature extraction methods with application to face recognition.,"* in In Systems, Man and Cybernetics, 2007. ISIC. IEEE International Conference on,IEEE, 2007., 2007, pp. 3627-3632.

Mr. Abhishek Tomar is a M. Tech Scholar of RCET, Bhilai (C.G) India. He did his Bachelor of Engineering in Information Technology from CSVTU university, Bhilai (C.G.), India.

Mrs. Minu Choudhary is working as Reader in Department of Computer Science & Engg. , RCET, Bhilai (CG) India. She has received her M. Tech in Computer Science and Engineering from CSVTU university, Bhilai (C.G). She has published many much research papers in national and international journals and conferences.

Mr. Amit Yerpude is working as Reader in Department of Computer Science & Engg. , RCET, Bhilai (CG) India. He has received his M. Tech in Computer Technology from CSVTU university; Bhilai (C.G).He has published many much research papers in national and international journals and conferences.