Segregation of Extended Features from Degraded Indian Manuscripts'Folios

Lalit Prakash Saxena

Department of Computer Science, University of Mumbai
lasaxmail@gmail.com

Abstract: Substantial portions of presently existing manuscripts exist in the form of folios, for instance, as degraded documents. Manuscriptfolios interpretation produced manually and recognition procedures contain computational errors. This paper uses an automatic approach to examine the selection and the effectiveness of searching techniques for possible speciousfeatures for recognition. The proposed method consists of two basic steps. In the first step, degradedfeatures in manuscript folios are located and computing operations are applied to create a collection of segregation features in the basic unit of the model. The second step uses 4-connectivity and 8-connectivity to generate additional typical features, identify appropriate matching positions, and determine the degree of relevance of retrieved folio images to the research request, based on asegregationmodel. The results obtained show its effectiveness and indicate an improvement over standard methods such as recognition systems without horizontal or vertical text and character segmentation.

Keywords: Manuscript images, Degradedfolios, Euler Number, Holes, Segregation model, 4-connectivity, 8-connectivity

1. Introduction

Through the recent decades, scientists have directed broad explorations on various aspects of manuscript images or folios handling. In investigation, considerable information is still stored in Ancient Indian Manuscripts, including birch, palm leaf, handmade paper and cloth. Despite all the research that has been done in manuscript image processing, several problems are still commonly encountered in this field. Script document images produced by scanning and recognition software contain segregation errors, and the rate of errors increases significantly with the degradations of the document image.

Segregation is the process of determining irrelevant attributes from a collection of available contents based on a feature inevitable for reduction. Research has been conducted on theinteraction between recognition and segregation since few decades and has consistently shown that the results of operations based on the particular reduction of irrelevant attributes are often of lower magnitude than expected [1]. For example, consider Figure 2. Here the extent of degradation reveals damages that had happened to them with span of time. Corners are almost gone, holes and stains disrupts the literature to be recognized. Therefore, it becomes more invulnerable to have the knowledge about degradation features accounted in manuscript folios.The quality of the original document can be a problem for the following reasons:

- Hand-written (having touching characters)
- Yellow papers (hand-made paper)
- Stains (black spots, greenish appearance)
- Holes in documents (leading to discontinuity of words)
- Insects affected (termite, silverfish attack mainly on surfaces)

2. Related work

Segregation of features serves to segregate large image databases and return attributes the system considers relevant to generate report. Reference [2] uses the approximate shape of

Received: August 20th, 2011. Accepted: Maret 15th, 2012

features inan image to refine the retrievalprocess; however, this approach cannot disambiguate recognition errors. Attributes segregation from manuscript images is difficult because manual reporting derive from reporting operations such as marking, measuring, and calculating. Previous studies have tried to reduce features through separation algorithm steps [3]. Most approaches to the separation algorithm of separating attributes make use of the information retrieval methods. Attributes are detected by searching the image for features that do not appear in an information retrieval method. This leads to many false alarms, since an information retrieval method cannot possibly cover everything [4]. Many studies in this area, shows three common mistakes – feature variation, algorithm evaluation, and report analysis – make up to 80–90% of all reporting alterations. During two decades a lot of work in the field of information retrieval had done, which conducted many experiments to study recognition accuracy and retrieval effectiveness from recognition-generated images [9]. They showed the effects of feature extraction on script images and feedback using the recognition model. An automatic process provides feedback that uses information derived from known relevant and non-relevant script images to reformulate features, but cannot be used to compensate for manual reporting caused by degraded folio images. A survey of image degradation models proposed in the existing literature can be found in [5]. Furthermore, as shown in [6], the segregation model improves the directional smoothing by synthesizing a script's content not only through a set of attributes but also by considering the importance of the features in folios and their specifics in the matrix. After creating a set of augmented and subjective features with the separation algorithm, segregation model is used to report relevant attributes and to evaluate segregation performance.

Finally, regarding the segregation model [7], is based on a recent work by in which different image processing algorithms were implemented with the goal of assessing criteria for feature selection using a variety of image potentials.

3. Model procedure for proposed approach

The Segregation model is shown in Figure 1. The approach is described by the following detailed stages.

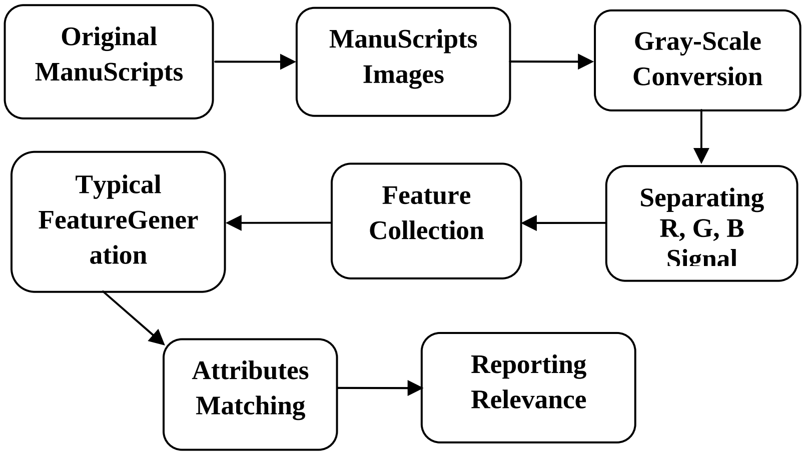

Figure 1. Segregation model based on attributes separation

In the first stage, we start from original manuscript folios (i.e. collecting, choosing and reporting). The first training sample contains document images and associated ground truth data (electronic image version) and is used for constructing features and reporting attributes.

The ground truths associated with document are the areas on each image and the corresponding attributes for all the featureregions.

- Lost area: any section vanishedarbitrarilyfrom the content of an original manuscript.
- Tears: each areamissing from manuscript linearly in a fashion.
- Worn out: parts detached when proper handling of manuscripts not taken.
- Holes: portions absent from manuscript folios making it see through.
- Others: the spread of stains (black or greenish) over each manuscript folio.

The second step uses features, attributes, data available in ground truth, andcomponents to create separable attributes, eliminate features, generate reports, identify appropriate reported features, and apply separation algorithm for separating and determining relevant attributes fromimages.

Finally, measure the performance of the segregation modeland compare with manual reporting to show the improvement in segregating attributes from images. The algorithm for segregation model is listed below:

```
Original Manuscript
I = Input Image;
G = Gray Scale Image of I;
B = Signal Vector of G (Separating Red, Green & Blue);
[LM, N] = Label Image B;
P = Feature Property (LM);
Q = [P. Feature];
[Z, X]=Size of Q;
X = X/4;
Q = Reshape Q with [4 X];
Display image B;
[Y, Collected Feature] = Size of Q;
Selected Line = 1;
For Feature = 1: Collected Line * Collected Column
    Place count on Q from feature == 1;
End
Display count (starting from 1);
For Feature = 1: Collected Line * Collected Column
    X=Q (: , count);
    C =X (2, 1);
    D =X (1, 1);
    E =X (3, 1) + D;
    F =X (4, 1) + C;
End
```

Details of the above algorithm related forFigure 1. arediscussed in the following sections.

4. Matching segregation features

Manual reportingis used to report the inhabited attributes in the manuscript images to perform feature separation. This manual reportingis time consuming and provokingsee Table 1 as an example. All measurements are in centimeters and shows the area covered in it.

*a,b are the sides of the folio, a-front side and b-back side; U, L are portion of folio, U-upper portion and L-lower portion;

Table 1.Showsan original report from University Library, Mumbai

Folio No.	Lost Area	Tears	Worn Out	Holes	Others

1a	0-0.1(U)
	0-0.2(L)
		10.2-10.4, 18.8
		21.8-22, 23.2-25.8, 26.5-28.1
			1.0-2, 9.3-10.3, 13.7-14.6, 20.9-21.7, 24.4-25.4, 28.2-28.8, 31.2-32.2	31.4, 38.6	0-45.5(L) all over Black stain

1b	__	__	6.5-8.3, 17.5-19		0-11(U)
2a	0-0.7(U) 34-till end(L)	19.5-20.7, 24.6-25.8, 26.6-28.9, 34-36.7, 37.8-39.9, 41-43.3	7.2-8.5, 17.8-19.4, 23.8-24.9, 30-31.3, 40.8-47.3(till end)	25.5, 25.6, 25.8, 31.4, 33, 37, 38.3, 43.9	0-10.5(U), 16-45.7(L) Black stain
2b	—		24.5-26, 32.5-34.2, 35.8-36.9		0-40(L), 0-45.4(U) till end Black stain
3a	0-0.6(U) 45-till end(L)	11, 11.3, 23.4-23.8, 25.6-26.3, 31-32.2, 35.2-35.4, 38, 41.2-41.4, 44.4-45.3	6.8-7.2, 19.2-20, 24.5-26.1, 27.4-28.5, 31-31.7, 32.5-34.9, 36-36.9, 38.5-47(till end)	24.9, 35.7, 36.8, 37.7	8.5-39.8(U), 0-45(L) till end Black stain
3b	—		25-25.5, 34.9-36.1		0-13(U), 36-47(L) till end Black stain

A. *Connectivity and connected components*

The concept of connectivity and connected components in an image arises as a set of connected pixels present in that image defining the concerned area [8]. The connected pixels(mainly neighborhood) of the image describe a metric arrangement of components respect to foreground and background.

Connected components:

- S= the set of object pixels
- S is a Connected Component if for each pixel pair (x1, y1) ∈S and (x2, y2) ∈S there is a path passing through X-neighbors in S. (X = 4, 8).
- S may contain several connected components.

For marking the connected components in an image a two pass Connected Component Algorithm is applied over it:

Pass 1: Scan the image pixels from left to right and fromtop to bottom.For every pixel P of value 1 (an object pixel), test top and leftneighbors (4-neighbor metric)

- If 2 of the neighbors equal 0: assign a new mark to P.
- If 1 of the neighbors equals 1: assign the neighbor's mark to P.
- If 2 of the neighbors equal1: assign the left neighbor's mark to Pand note equivalence between 2 neighbor's marks.

Pass 2: Divide all marks in to equivalence classes (marks of neighboringpixelsare considered equivalent).

Replace each mark with the number of its equivalence class.

B. Euler number

Euler number is defined as the difference between number of connected components and number of holes in a binary image [8]. Hence if an image has C connected components and H number of holes, the Euler number E of the image can be defined as:

E = C – H;Euler Number = the number of objects minus the number of holes.

Separation algorithm segregates holes returns the Euler number for the binary image. Euler number is a scalar whose value is the total number of objects in the image minus the total number of holes in those objects. It can have a value of either 4 or 8 as an argument, where 4 specifies 4-connected objects and 8 specifies 8-connected objects; if the argument is omitted, it defaults to 8.Algorithm computes the Euler number by considering patterns of convexity and concavity in local 2-by-2 neighborhoods.

Calculating the euler number

Euler = S – S'
S= object pixels
S'= all other pixels
Background= connected components of S'that touch the edge of the image.
Hole= connected components of S'that is not in the background.
Simply Connected Component = a component without holes.

Separation Algorithm estimates the area of all of the on pixels in an image by summing the areas of each pixel in the image. The area of an individual pixel is determined by looking at its 2-by-2 neighborhood. There are six different patterns distinguished, each representing a different area:

- Patterns with zero on pixels (area = 0)
- Patterns with one on pixel (area = 1/4)
- Patterns with two adjacent on pixels (area = 1/2)
- Patterns with two diagonal on pixels (area = 3/4)
- Patterns with three on pixels (area = 7/8)
- Patterns with all four on pixels (area = 1)

5. Segregation process

Attributes segregation is about finding the significantattributesin a folio image, with features being one of its basic tools. However, exact attribute separation is not good enough for folio image separated because a feature,when recognized incorrectly in the image, can no longer be reported. The expanded, reportedfeature is separated against the availablefolio of images obtained by the $M \times N$ image matrix, where M is the number of attributes in the image and N is the number of unique features in the image. The similarity of each attribute is calculated in the subset of image, and the system generates an organizedreportfor it. Finally, measuring the performance of the segregation modeland compare it with manual reporting method.

A. Attributes recognition and selection

For every attribute, the features generated by relieving all attribute contained in the folio image with their corresponding separation algorithm. Let us first give an example of the feature separation. Suppose that a manual report contains anattribute "Lost area". It is statistically uncertain because manually it is not possible to measure fine edges in a manuscript folio. This can, however, also help in identifyingthe images in which "Lost area" has been mistakenlyrecognized as "holes".

B. Boundary and area calculations

Boundary and area measurements are meaningful only for binary images [8]. Consider a discrete binary image containing one or more features, where P (j, k) = 1 if a pixel is part of the object and P (j, k) = 0 for all non-object or background pixels. The area of each feature within the image is simply the count of the number of pixels in the object for which P (j, k) = 1. As an example, for a 2×2 pixel square, the feature area is A_F= 4 and the featureboundary is B_F = 8. A feature formed of three diagonally connected pixels possesses A_F = 3 and B_F = 12. The enclosed area of a feature is defined to be the total number of pixels for which P(j, k) = 0 or 1 within the outer boundary $B_O$of the object. Then, proceeding in a clockwise direction around the boundary, a crack code B(c) is generated for each side p of the object boundary such that B(c)= 0, 1, 2, 3 for directional angles 0, 90, 180, 270°, respectively.

C. Distance algorithm

Distances is determined in the following manner for a pixel in an image-

Two grid point: P = (x,y) and Q = (u,v)

- Euclidean Distance
  $d_e(P,Q) = \sqrt{\{(x-u)^2 + (y-v)^2\}}$
- City Block Distance
  $d_4(P,Q) = |(x-u)|+ |(y-v)|$
- Chessboard Distance
  $d_8(P,Q) = \max(|(x-u)|, |(y-v)|)$

Distances de, d8, d4 are all metrics:

- Distance metric: $d(P,Q) \geq 0$
- Positive: $d(P,Q) = 0$ iff $P=Q$
- Symmetric: $d(P,Q) = d(Q,P)$
- Triangular inequality: $d(P,Q) \leq d(P,R) + d(R,Q)$

For each pixel calculate the $d_4$or $d_8$distance from a pixel in set S

2 passes:

Pass 1: scan image left-to-right and top-to-bottom

Pass 2: scan image right-to-left and bottom-to-top.

For each pixel Pmark as follows:

D. Performance measures

Performance is determined by the separation of randomly selected attributes. The lists of relevant attributes ofscript images on the basis of original manuscripts are compared with those obtained by manual reporting. The evaluation of these two methods is based on the separation effectiveness using average values ofthe assessedseparation and precision, which are calculated from the following equations:

*D.1. Separation***:**is a measure of the ability of the model to separate all relevant attributes. It is calculated as

$$\text{Separation} = \frac{\text{total of relevant attributes separated}}{\text{total number of relevant attributes}}$$

D.2. Precision: is a measure of the ability of the modelto separate only relevant attributes. It is calculated as

$$\text{Precision} = \frac{\text{total of relevant attributes separated}}{\text{total number of attributes separated}}$$

6. Experimental results

A. Data collection

Starting with training data (TD) which was used to construct attributes; folio images were the technical images database of the segregation model with 22 script images from a variety of manuscripts covering diverse degradation and reporting problems, the average number per image was 8 attributes. Manual reporting (MR) was done on 58 manuscripts with more than 100 folio image search having average 6 attributes per script folio. For feature collection, 10 features randomly selected from the content of folio image; each feature contained on an average 0.4cm^2 areas. The reports generated by different systems were then compared with the supposed relevant attributes to determine, for each feature, whether or not they were relevant.

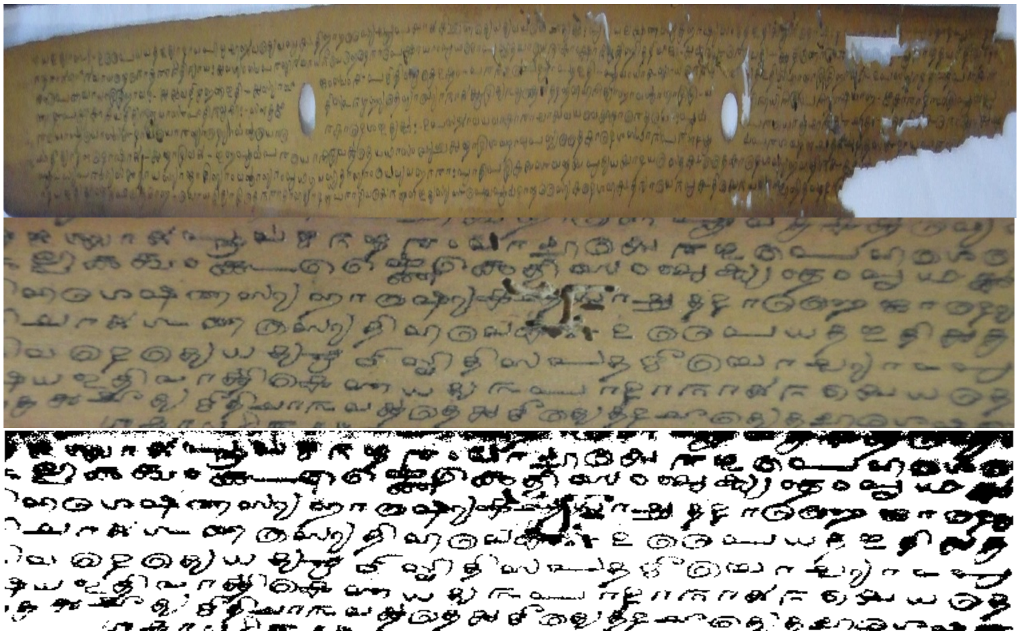

Figure 2.ashows an original manuscript folio image and types of degradations, b is portion of folio and c is subsequent separation result.

B. Manual segregation

B.1. Segregating phase:

The result obtained using separation algorithm is presented in Table 2. The original folio images had 1035 attributes, which explain the higher number of features present in the script image than the original folio: 958 attributes present in original manuscript folio while the script image extracted 831 attributes. Only 831 attributes out of 1035 were correctly extracted. The separation algorithm matched 767 attributes as specific attributes and used them to generate more features. The algorithm generated specific attributes and generated segregation specific attributes and generated separation rules. Table 2.shows the percentage of specific of the original script folios and script images.

Table 2. Attribute separation and manual errors on folio images. 22 script images obtained from 12 manuscript folios

	Original manuscript folio	Script image
Number of attributes	1035	1035
Manual segregation	958	831
Number of specific attributes	767	583
% of specific attributes	74.10	56.32

B.2. Test phase:

Table 3, Table 4 show the results of the recognition segregation of the manual reporting and segregation model. Figure 3shows the decrease in the performance of the segregation of the script images and observes that degradation factor η, affect segregation accuracy. The separation percentages of various attributes, but the performance falls as the specific features for "lost area" to "others".

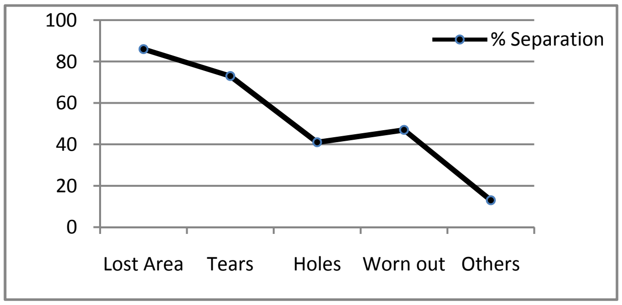

Figure 3.Separation percentage of segregated attributes

Table 3. Attribute separation and manual errors on folio images

Specific attributes		Worn out			Others	
Degradation factor, η	0.1	0.01	0.001	0.1	0.01	0.001
Number of attributes		1035			1035	
Manual segregation	565	578	589	259	247	252
Specific attributes	431	419	459	87	81	83
% of specific attributes	41.64	40.48	44.35	8.41	7.80	8.02

22 script images interpreted by degradation factor with three standard deviations and with two specific attributes

Table 4. Separation-precision results on script images

	Low-separate	Mid-separate	High-separate	Average precision
Manual segregation	96.83-98.89	82.66-95.92	26.63-73.43	75.06%
Best average attributes	95.45-99.5	39.98-88-45	0-31.66	59.16%
This method	96.12-98-72	84.98-93.27	37.39-83.44	92.13%

C. Segregation effectiveness

The separation-precisiongraph is the most regularly used method for comparing methods. The plots of different runs can be superimposed on the same graph to determine which method is superior. Comparisons are best made in three different separate ranges: *0–0.2, 0.2–0.8,* and *0.8–1*. These ranges characterize low-separate, mid-separate, and high-separate performance, respectively.

*C.1. On holes***:** In Figure 6, Table 4 for the MR method without expanded features, the average precision is between 96.83 and 98.89% for the low-separate, between 82.66 and 95.92% for the middle- separate and between 26.63 and 73.43% for the high-separate performance.

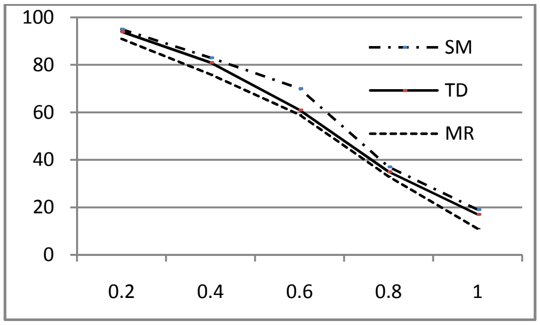

Figure 6. Separation-precision averages for wornout

C.2. *On tears, lost area, worn out and others:* The SM and MR collections used to obtain the results presented in Figure 4, Figure 5, Figure 7, Table 3, Table 4 show the same tendency as that of the TD, except for the best average attributes approach, with a degradation factor, which is able to perform well in the high precision field. For SM, the average precision for all specific features (averaged over features) is 92.13%, but 85.33% for the TD without expanded query, and it does not exceed 75.06% for MR. For the "worn-out" area, the results obtained in Table 3. Shows the best overall precision and concurs with the separation of the best average attributes. The explanation for this is obtained from undetermined features in the script image, such as "tears" or "lost area" and which are inexplicitlyreported. We can see in Figure 4, Figure 5, Figure 7that the precision is maintained with an upper limit of 65% for separation lower than 50%. The average precision for all specific attribute over all features is 63% for Segregation method, but decreases to 59% for TD and to 60% for MR without expanded features. However, the rate decreases to 50% for TD. An indication of the results obtained for "worn-out" can be seen in Figure 7, which shows a drop in the precision rate. The problem with the MR approachis the fast drop in precision when the separate variate is low.

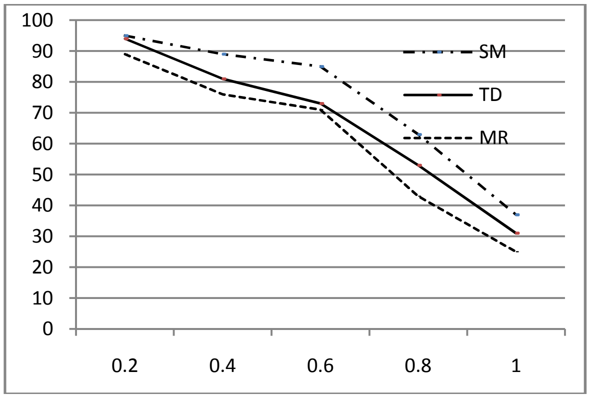

Figure 4. Separation-precision averages for lost area

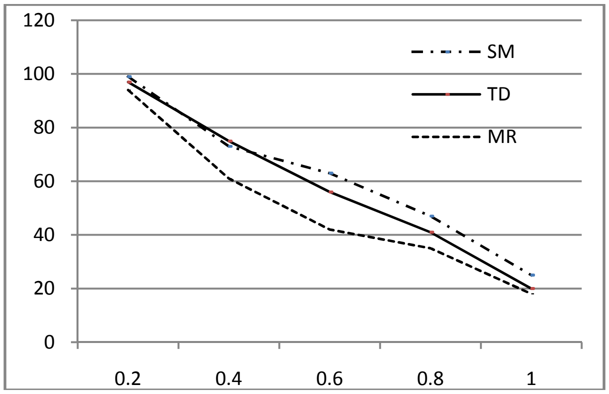

Figure 5. Separation-precision averages for tears

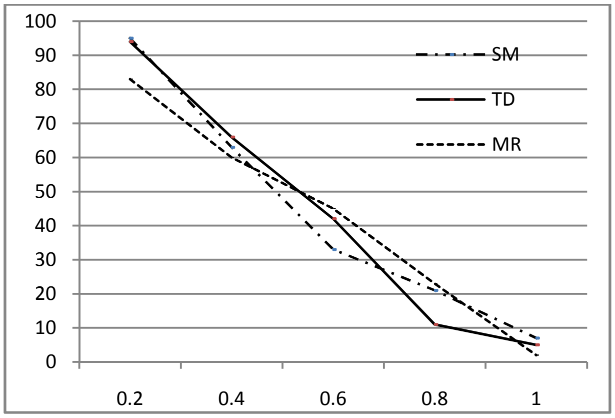

Figure 7. Separation-precision averages for holes

7. Conclusions

This work presents an approach to processing segregated attributes contained in manuscript images and for performing effective separation algorithm. Manually preparing reports does not have significant details about the segregation attributes. The proposed method collects frequent segregation attributes and separation algorithm that can be used to extend features and to improve segregation performance. Furthermore, investigating the attributes segregation of poor-quality manuscripts is important for folio images generated from archives of originals created before the dawn of the digital age. Further research is currently being undertaken to outperform our approach. The aim is to investigate the ability of this approach to improve separability.

8. Acknowledgement

We gratefully acknowledge Dr. Anjali Kade for providing and allowing us to take photograph of Original Manuscripts used in this experiment; Dr. VandanaLele for detailing about the segregation attributes and her colleague, Dr. Sujata for detailing the manual reporting system of attributes separation.

9. References

[1] S. Crisan, I. G. Tarnovan, T. Eduard andC. Vein, "Pattern Recognition: Image enhancement and feature extraction algorithms,"*Journal of Computational Physics*, 2008.

[2] H. Motameni, M. Norouzi, M. Jahandar and A. Hatami,"Labeling Method in Steganography," *World Academy of Science, Engineering and Technology*, vol. 30, 2007.

[3] T. Nawaz, K. A. Qazi and M. I. Ashraf,"Performance Evaluation of Noise Removal Algorithms for Scanned Images,"*International Journal of Computer Science and Security*, vol. (3) pp. 226-229, 2009.

[4] G. Phani, D. Maruti, V. Borker and J.Sivaswamy,"Impulse Noise Removal from Color Images with Hopfield Neural Network and Improved Vector Median Filter," in Proc. Sixth Indian Conference on Computer Vision, Graphics & Image Processing, *IEEE computer society*, pp. 17-24, 2008.

[5] M. Saeidi, K. Saeidi and M. Khaleghi, "Noise Reduction in Image Sequences using an Effective Fuzzy Algorithm,"*World Academy of Science, Engineering and Technology*, vol. 43, pp. 351-356, 2008.

[6] H. Sahoolizadeh, R. Rajabioun and M. Zeinali, "A Fourier Extension Based Algorithm for Impulse Noise Removal," in Proc. the World Congress on Engineering, 2008, paper 1, p. 2.

[7] O. P. Verma, M. Hanmandlu, A. Parihar and V. K. Madasu,"Fuzzy Filters for Noise Reduction in Color Images,"in *Proc. ICGST-GVIP'09*, 2009, paper 9.5, p. 29.

[8] R. C. Gonzalez and R. E. Woods,*Digital Image Processing*, 2nd Ed.: New Jersey:Prentice Hall Upper Saddle River, 2002.

[9] M. Wu and J. Farquhar,“A Subspace Kernel for Nonlinear Feature Extraction,” in *Proc.International Joint Conference on Artificial Intelligence*, 2007, p. 1125.

Lalit Prakash Saxena was born in Obra, Sonebhadra in 1984. He graduated in Physics and Mathematics from Government Post–Graduate College, Obra in 2006. He has done his Masters in Computer Applications from Bundelkhand University, Jhansi in 2009. He is currently doing PhD in Computer Science from Department of Computer Science, University of Mumbai, Mumbai. His main research interest is in Image Processing, Pattern Recognition, Document Image Analysis and Manuscript Image enhancement and Script Classification.