Digitization and Recognition of Historical Kannada Handwritten Manuscripts using Text Line Segmentation with LBP Features

[1] Parashuram Bannigidad, [2] Chandrashekar Gudada

[1] Assistant Professor, [2] Research Scholar
[1, 2] Department of Computer Science,
[1, 2] Rani Channamma University, Belagavi, KA, India

Abstract : The inscriptions or Epigraphical manuscripts where composed on various material, for example, walls on the caves, stone carving, palm leaf, metal plates and paper are the resources and cultural heritage of our country; our focus is to reproduce the cultural significance of the Kannada Language and its traditional writing through the historical manuscripts. The majority of the assets are in the degraded state, the degraded manuscripts are affected by many factors like, weather condition, ink bleed through and quality of the writing material. The offline handwritten text recognition is one of the most challenging tasks in document image analysis. In the present era of digital, it's our fundamental duty to protect the resources of our Indian culture and heritage by digitizing the manuscripts which are losing its originality and status. In this paper, we are trying to identify and recognize the historical Kannada handwritten scripts of various dynasties; namely, Vijayanagara dynasty (1460 AD), Mysore Wadiyar dynasty (1936 AD), Vijayanagara dynasty (1400 AD) and Hoysala dynasty (1340 AD) by using seam carving line segmentation method by extracting LBP features. For recognition and classification; the LDA, K-NN and SVM classifiers are used. The average classification accuracy for different dynasties is computed. The computed results are yielded an overall accuracy of 94.2% for LDA, 94.9% for K-NN and 96.4% for SVM is achieved, based on the experimented results, the SVM classifier has yielded the higher classification accuracy comparatively LDA and K-NN classifiers for historical Kannada handwritten scripts. The experimental outcomes are verified by language experts and Epigraphists.

***IndexTerms* - Document image analysis, Historical documents, Handwritten script, Seam carving, Line segmentation, Kannada, LBP, LDA, K-NN, SVM.**

I. INTRODUCTION

India is the oldest and ancient civilized country in the world, its civilization evolved before 7000 BCE with spiritual and astrological knowledge. This knowledge of information is stored and kept preserved in the form of historical inscriptions and epigraphical manuscripts in the manuscript resource centres, manuscript conservation centres, manuscript partner centres, gurukula and monasteries. Due to the negligence in maintaining, these scripts are in the form of degradedness. Hence, the digitization of these degraded documents is an important task to restore the deciphering inscriptions. Particularly, in the state of Karnataka, many dynasties have ruled and contributed their knowledge to the Indian civilization. In this work, we experiment with the available historical Kannada handwritten manuscripts (Paper inscription) of various dynasties, namely; Vijayanagara dynasty (1460 AD), Mysore Wadiyar dynasty (1936 AD), Vijayanagara dynasty (1400 AD) and Hoysala dynasty (1340 AD), we have collected these inscriptions from different manuscript preservation centers by capturing high resolution digital camera and are stored them in the digital form for further identification and recognition of historical Kannada handwritten manuscripts.

Very few authors have contributed to this area in the literature; The text line extraction using seam carving for colour and grayscale historical manuscript was proposed by Nikolaos et.al.[4]. The content-aware image resizing using Seam carving method has been investigated by Avidan et.al.[5]. The significance of text line segmentation in handwritten text recognition was presented by Romero et.al.[6]. The writer identification using sparse radial sampling LBP features was proposed by Nicolaou et.al.[7]. The English Word spotting in handwritten historical document by using LBP features has been proposed by Dey et.al.[8]. The LBP based script identification based on text line was investigated by Ferrer et.al.[9]. Ghosh et.al.[10] have proposed an algorithm for separation of text / non-text from handwritten document images using LBP features. The evaluation for historical document image analysis using texture features was carried out by Mehri et.al.[11]. Laurence et.al.[12] has done a survey on text line segmentation of historical documents. Text line segmentation for gray scale historical document images was proposed by Asi et.al.[13]. Parashuram and Chandrashekar [14, 15] have proposed an algorithm of image enhancement and binarization method for degraded historical Kannada handwritten manuscript document images. The classification of historical Kannada handwritten manuscript images based on their age-type using LBP features was proposed by Parashuram and Chandrashekar [16].

II. PROPOSED METHOD

The main objective of the proposed method is to digitize, identify and recognize the historical Kannada handwritten manuscripts based on their different age-type using seam carving line segmentation method by extracting LBP features and LDA, K-NN and SVM classifiers. The detailed approach of the proposed method is discussed in the form of algorithm, which is described below:

Algorithm for recognition of Historical Kannada Handwritten manuscript

```
1.  Input Camera capture historical Kannada handwritten manuscript of different age-types: namely Hoysala,
    Vijayanagara and Mysore dynasties.
2.  Apply Improved seam carving text line segmentation method for line Extraction from historical manuscript:
    2.1.  convert the given original colour image to grayscale image
    2.2.  compute edge image using the Sobel edge detector
    2.3.  medial seam computation with a projection profile matching
          2.3.1. Compute horizontal projection profiles of all edge image slices and find their local maxima
          2.3.2.  Match local maxima of the projection profiles between two consecutive image slices
          2.3.3.  Remove lines that start from some intermediate column of the image
          2.3.4.  Extend the small lines towards the end column of the image
    2.4.  Separating seam computation with constrained seam carving
          2.5.1.  apply constrained seam carving for each pair of text lines
          2.5.2.  compute minimum energy separating seam using dynamic programming
          2.5.3.  overlay separating seams on the original image
                  2.5.3.1. Compute the coordinate values of the overlay separating seam and store them in temp
                  2.5.3.2. Concatenate the present coordinate values with temp
          2.5.4.  Using coordinate value, extract the region of interest by roipoly() function
          2.5.5.  Apply the image enhancement technique for binarization to the extracted region of interest (text line)
    2.5.  Original image overlaid with both types of seams
    2.6.  Apply the skew correction to the segmented text line
3.  Combination of Local Otsu and Global Otsu method is applied to each individual text line for binarizing the images on
    step 2
4.  Apply size normalization to each individual text line on step 3
5.  Extract LBP features of size normalized individual text line of different dynasties, namely; Hoysala, Vijayanagara and
    Mysore dynasties and store them as a knowledge base
6.  Apply the classification techniques, namely; LDA classifier, K-nearest neighbour classifier and SVM classifier to
    classify and recognize the historical Kannada handwritten manuscripts, whether they belong to the Hoysala dynasty or
    Vijayanagara dynasty or Mysore dynasty?
```

The detailed approach of the proposed method is given in the Figure 1.

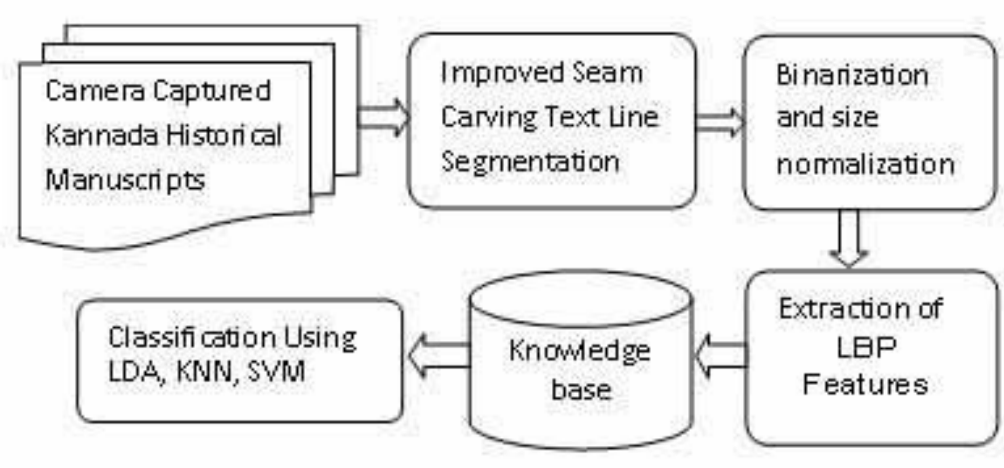

Figure 1. The detailed approach of the proposed method

III. Experimental results and discussion

We have considered the datasets of different dynasties, namely; Vijayanagara(1400 AD and 1460 AD), Hoysala(1340 AD) and Mysore Wadiyar(1936 AD) (described in Sect. II) for experimentation. The experimentation is done with Intel Core i5 system using Matlab R2018b. Input the camera captured historical Kannada handwritten manuscript document images (Figure 2a) for extraction of text line segmentation. To extract text line segmentation, we applied the improved seam carving method which includes the computation of medial seam (Figure 2b), Separating seam computation with constrained seam carving based on medial seam (Figure 2c), overlaid with both type of seam (Figure 2d), region of interest i.e., text line is extracted based on the overlaid seam using roipoly() function (Figure 2e). And then apply the image enhancement method for binarization, restoration and size normalization to each individual text line (Figure 2f). Extracted the LBP features for all the text lines and store them as a knowledge base. Finally, apply classification techniques; i.e., LDA classifier, K-NN classifier and SVM classifier for classification and recognition of the historical Kannada handwritten manuscripts based on their age-type. The other sample images of the proposed algorithm used for other dynasties namely, Mysore Wadiyar(1936 AD) dynasty, Vijayanagara(1400 AD) dynasty and Hoysala(1340 AD) dynasty, which are shown in Figure 3, Figure 4 and Figure 5, respectively. The average classification accuracy of the proposed method is given in the Table 3. Initially, we have extracted the LBP features with 59 features using k-fold experiment and these results are given in the Table 1. Further, to improve the results we have reduced the LBP features with 19 features and certainly results are improved. The results of these reduced features with k-fold experiments using LDA, K-NN and SVM classifiers are given in the Table 2. As per the results, it is observed that LBP features with 19 features are given better results comparatively 59 features. Hence, we propose only 19 features by reduced features from 59 and overall accuracy is calculated and represented only based on 19 LBP features.

The classification accuracy for different dynasties based on their age-type is represents that the LDA classifier has yielded 94.2%, K-NN classifier has 94.9% and SVM classifier has 96.4%. Based on the experimentation, the result shows that the SVM classifier has got a good classification performance comparatively LDA classifier and K-NN classifier for historical Kannada handwritten manuscript images. The results of the confusion table of various classification techniques are given in the Table 4, 5, 6, which indicates better recognition rates towards historical Kannada handwritten manuscript images.

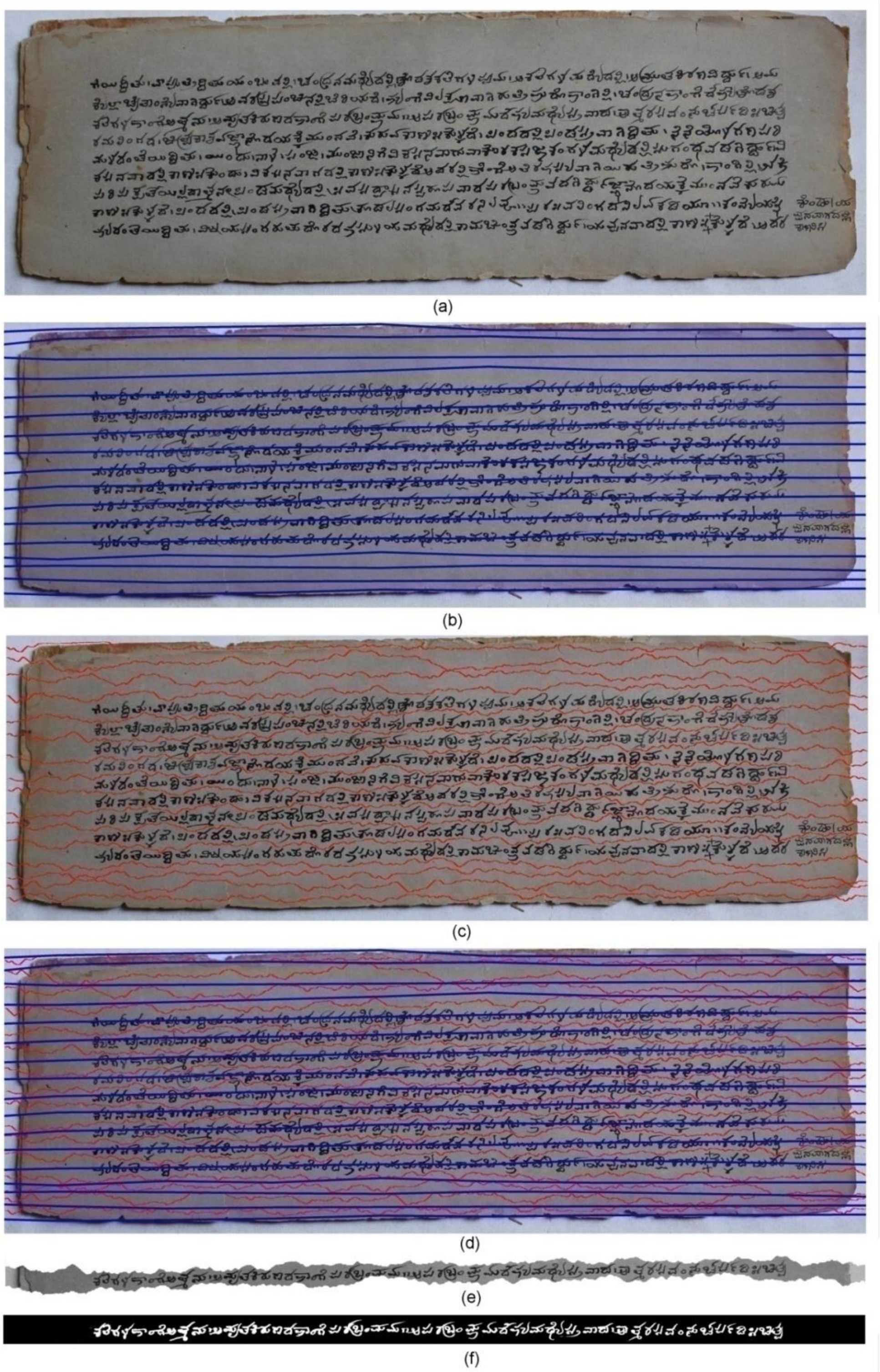

Figure 2. Sample image of the Vijayanagara dynasty (1460AD) (a) Original camera captured image (b) Medial seam computed image (c) Separated seam computed image with constrained seam carving based on medial seam (d) Overlaid image of both type of seam i. e. medial seam and carved seam (e) Text line is extracted based on the carved seam (f) Enhanced and size normalized image on (e)

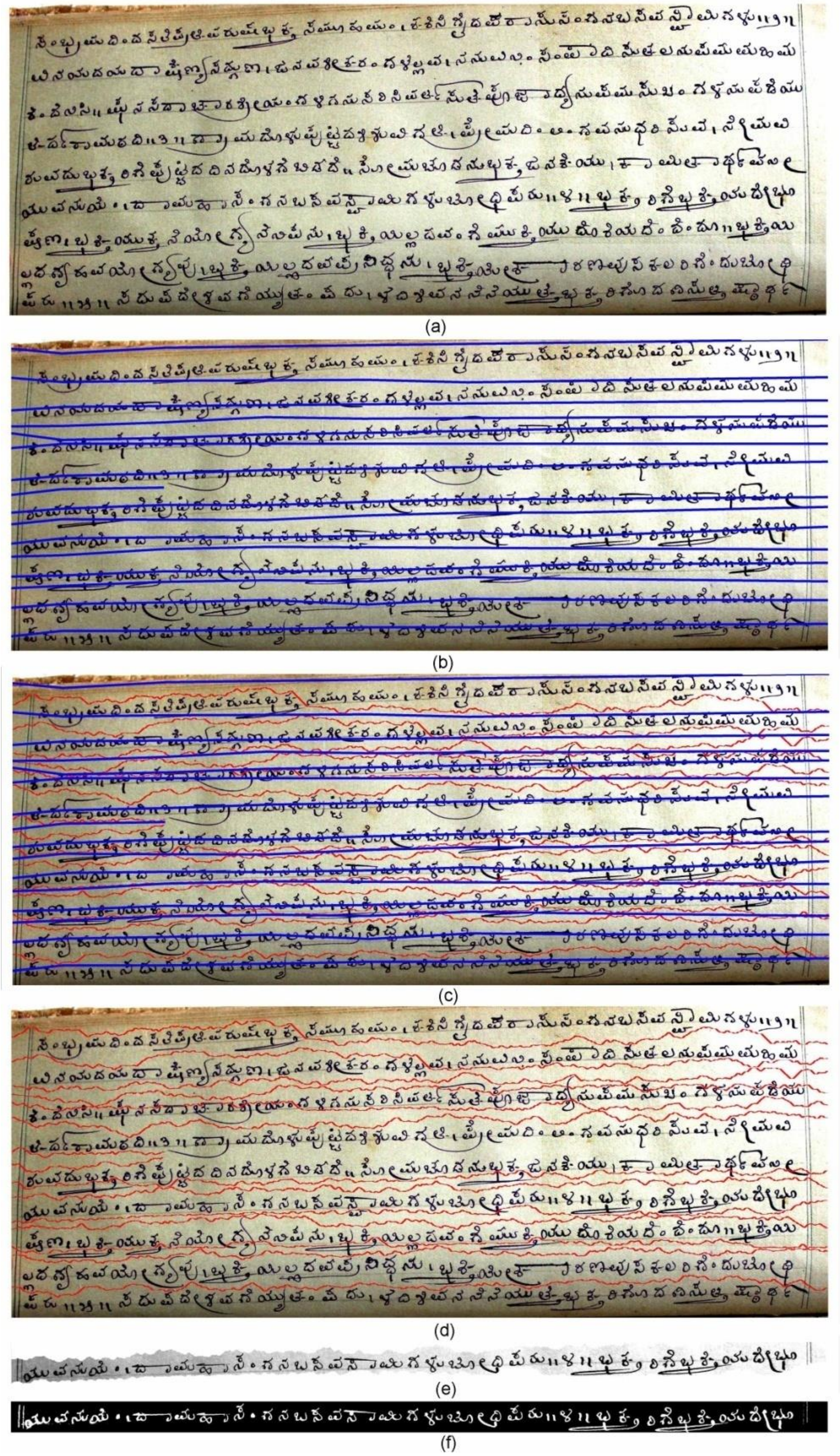

Figure 3. Sample image of Mysoure wodeyar dynasty (1936AD) (a) Original camera captured image (b) Medial seam computed image (c) Separated seam computed image with constrained seam carving based on medial seam (d) Overlaid image of both type of seam i. e. medial seam and carved seam (e) Text line is extracted based on the carved seam (f) Enhanced and size normalized image on (e)

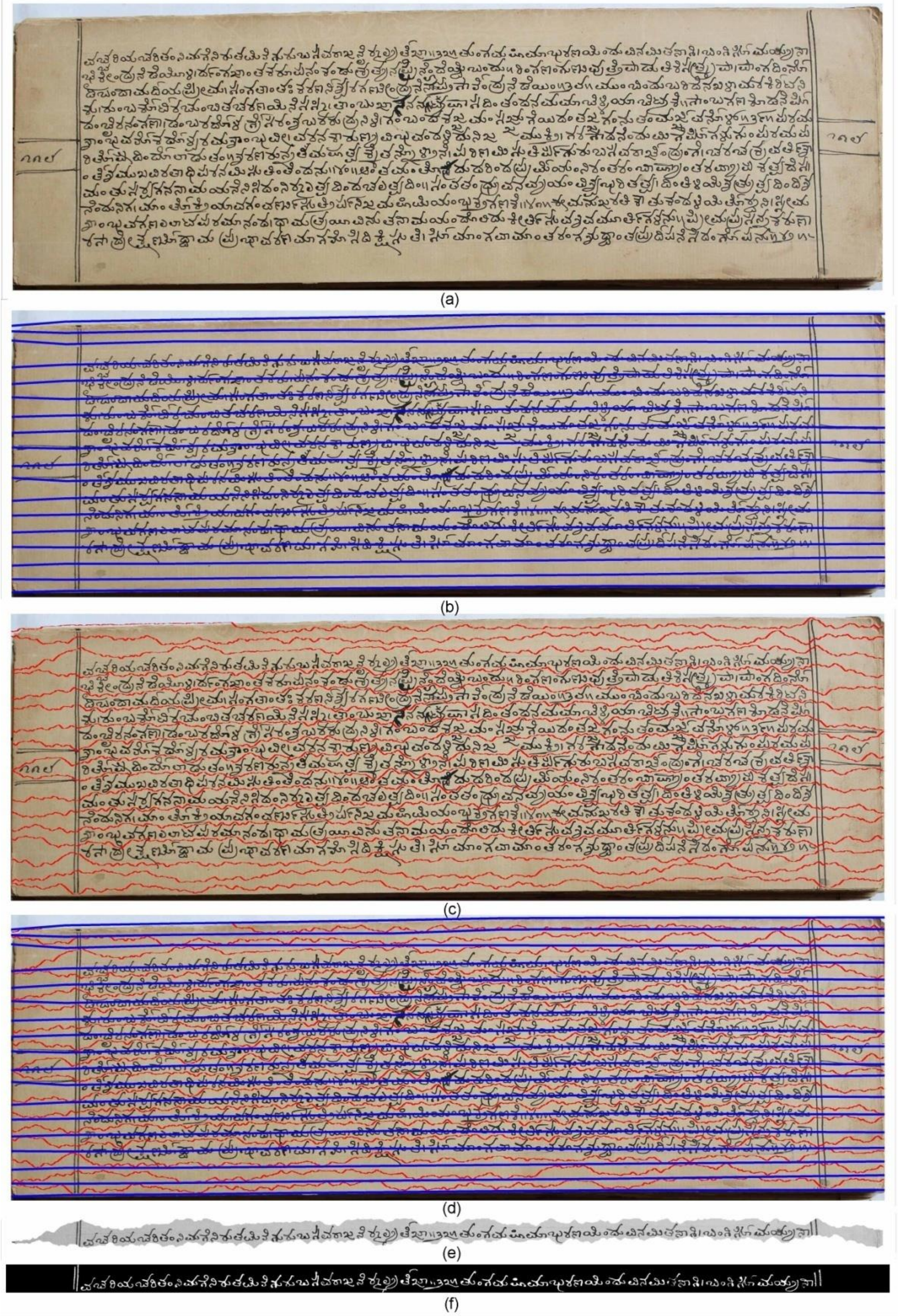

Figure 4. Sample image of Vijayanagara dynasty (1400AD) (a) Original camera captured image (b) Medial seam computed image (c) Separated seam computed image with constrained seam carving based on medial seam (d) Overlaid image of both type of seam i. e. medial seam and carved seam (e) Text line is extracted based on the carved seam (f) Enhanced and size normalized image on (e)

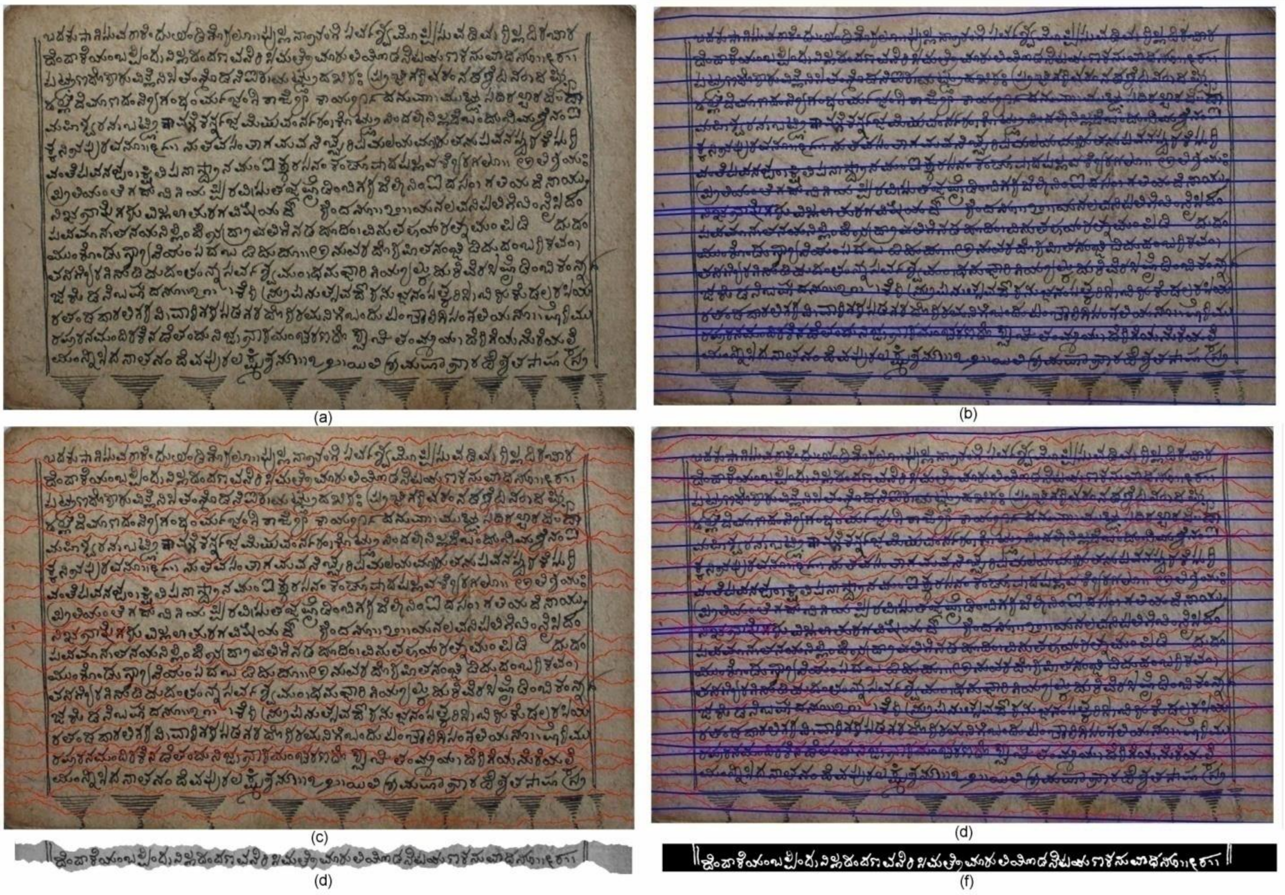

Figure 5. Sample image of Hoysala dynasty (1340AD) (a) Original camera captured image (b) Medial seam computed image (c) Separated seam computed image with constrained seam carving based on medial seam (d) Overlaid image of both type of seam i. e. medial seam and carved seam (e) Text line is extracted based on the carved seam (f) Enhanced and size normalized image on (e)

Table 1.Classification accuracy of LBP with 59 Features for different k-fold experimentation

Classifiers	5 Fold	4 Fold	3 Fold	2 Fold
LDA	NA	NA	NA	NA
KNN	94.9	94.5	94.8	93.9
SVM	96.4	96.1	95.4	95.7

Table 2.Classification accuracy of LBP with 19 Features for different k-fold experimentation

Classifiers	5 Fold	4 Fold	3 Fold	2 Fold
LDA	94.2	94.1	93.8	93.9
KNN	94.9	94.9	93.5	93.7
SVM	96.4	96.1	94.5	95.9

Table 3.The average classification accuracy of proposed method with LDA, K-NN and SVM classifiers

Dynasties	LDA		K-NN		SVM	
	Recognition Rate	Error Rate	Recognition Rate	Error Rate	Recognition Rate	Error Rate
Vijayanagara(1460)	96	4	98	2	97	3
Mysoure Wodeyar(1936)	89	11	87	13	92	8
Vijayanagara(1400)	97	3	97	3	98	2
Hoysala(1340)	93	7	95	5	97	3
Average accuracy	**94.2%**		**94.9%**		**96.4%**	

Table 4.Confusion table for LDA classifier

Dynasties	Vijayanagara (1460)	Mysoure Wodeyar (1936)	Vijayanagara (1400)	Hoysala (1340)	Total No. of Text line
Vijayanagara (1460)	**240**	3	6	1	250
Mysoure Wodeyar (1936)	3	**194**	21	1	219
Vijayanagara (1400)	1	9	**395**	1	406
Hoysala (1340)	1	7	16	**310**	334

Table 5.Confusion table for K-NN classifier

Dynasties	Vijayanagara (1460)	Mysoure Wodeyar (1936)	Vijayanagara (1400)	Hoysala (1340)	Total No. of Text line
Vijayanagara (1460)	**245**	2	3	0	250
Mysoure Wodeyar (1936)	6	**190**	18	5	219
Vijayanagara (1400)	2	9	**394**	1	406
Hoysala (1340)	3	5	8	**318**	334

Table 6.Confusion table for SVM classifier

Dynasties	Vijayanagara (1460)	Mysoure Wodeyar (1936)	Vijayanagara (1400)	Hoysala (1340)	Total No. of Text line
Vijayanagara (1460)	**245**		5		250
Mysoure Wodeyar (1936)	3	**198**	11	9	219
Vijayanagara (1400)	3	2	**393**	8	406
Hoysala (1340)		10	5	**319**	334

IV. CONCLUSION

In this paper, we have proposed an algorithm to identify and recognize the historical Kannada handwritten scripts of different dynasties; namely, Vijayanagara dynasty (1460 AD), Mysore Wadiyar dynasty (1936 AD), Vijayanagara dynasty (1400 AD) and Hoysala dynasty (1340 AD) by using seam carving line segmentation method by extracting LBP features. For recognition and classification; the LDA, K-NN and SVM classifiers are used. The average classification accuracy for different dynasties is computed. The computed results are yielded an overall accuracy of 94.2% for LDA, 94.9% for K-NN and 96.4% for SVM is achieved, based on the experimented result the SVM classifier is yielded the higher classification accuracy comparatively LDA and K-NN classifiers for historical Kannada handwritten scripts. The experimental outcomes are verified by language experts and Epigraphists, which shows the robustness of the proposed method. The same algorithm may be used for other dynasties with different feature sets, which will be done in the future work.

REFERENCES

[1] Manjunath, M.G., Devarajaswamy G.K., “Kannada Lipiya Vikasa”, Published by Jagadhguru Sri Madhvacharya Trust, Sri Raghavendra Swami Matta, Mantralaya

[2] Narasimha Murthy, A.V., “Kannada Lipiya Ugama Mattu Vikasa”, Kannada Adhyayana Samsthe, Mysore University, Mysore (1968)

[3] Reddy D., “Lipiya HuttuMattu Belavanige—Origin and Evolution of Script”, Kannada Pustaka Pradhikara (Kannada Book Authority), Bangalore

[4] Nikolaos Arvanitopoulos and Sabine Susstrunk, “Seam Carving for Text Line Extraction on Color and Grayscale Historical Manuscripts” 14th IEEE International Conference on Frontiers in Handwriting Recognition, DOI: 10.1109/ICFHR.2014.127, pp.726-731.

[5] S. Avidan, A. Shamir, “Seam Carving for Content-Aware Image Resizing,” ACM Transactions on Graphics, vol. 26, no. 3,p. 10, 2007.

[6] Veronica Romero, Joan Andreu Sanchez, Vicente Bosch, Katrien Depuydt and Jesse de Does, “Influence of Text Line Segmentation in Handwritten Text Recognition” 13th IEEE International Conference on Document Analysis and Recognition (ICDAR), 2015, 978-1-4799-1805-8/15, pp.536-540

[7] Anguelos Nicolaou, Andrew D. Bagdanov, Marcus Liwickiy, and Dimosthenis Karatzas, "Sparse Radial Sampling LBP for Writer Identification", 2015 13th International Conference on Document Analysis and Recognition (ICDAR), doi: 10.1109/ICDAR.2015.7333855, pp.716-720.

[8] Sounak Dey, Anguelos Nicolaou, Josep Llados , and Umapada Pal, "Local Binary Pattern for Word Spotting in Handwritten Historical Document", Structural, Syntactic, and Statistical Pattern Recognition. 2016. Lecture Notes in Computer Science, vol 10029. Springer, Cham, pp.574-583.

[9] Miguel A. Ferrer, Aythami Morales and Umapada Pal, "LBP Based Line-wise Script Identification", 2013 12th International Conference on Document Analysis and Recognition, doi:10.1109/ICDAR.2013.81, pp.369-373.

[10] Sourav Ghosh, Dibyadwati Lahiri, Showmik Bhowmik, Ergina Kavallieratou and Ram Sarkar, "Text/Non-Text Separation from Handwritten Document Images Using LBP Based Features: An Empirical Study", J. Imaging 2018, 4, 57; doi:10.3390/jimaging4040057, pp.1-15.

[11] Maroua Mehri, Pierre Héroux, Petra Gomez-Krämer, Rémy Mullot, "Texture feature benchmarking and evaluation for historical document image analysis", International Journal on Document Analysis and Recognition, Springer Verlag, 2017, pp.1-35. doi:10.1007/s10032-016-0278-y.

[12] Laurence Likforman-Sulem, Abderrazak Zahour and Bruno Taconet, "Text line segmentation of historical documents: a survey" IJDAR (2007) pp:123–138, DOI 10.1007/s10032-006-0023-z

[13] Abedelkadir Asi, Raid Saabni and Jihad El-Sana, "Text Line Segmentation for Gray Scale Historical Document Images", HIP '11 Proceedings of the 2011 Workshop on Historical Document Imaging and Processing, DOI: 10.1145/2037342.2037362, pp. 120-126, 2011.

[14] Parashuram Bannigidad and Chandrashekar Gudada, "Restoration of Degraded Historical Kannada Handwritten Document Images using Image Enhancement Techniques", International Conference on Soft Computing and Pattern Recognition (SoCPaR 2016), 2016. pp. 498-508

[15] Parashuram Bannigidad and Chandrashekar Gudada, "Restoration of Degraded Kannada Handwritten Paper Inscriptions (Hastaprati) using Image Enhancement Techniques", IEEE International Conference on Computer Communication and Informatics (ICCCI -2017), 2017. Pp.1-6.

[16] Parashuram Bannigidad, Chandrashekar Gudada, "Identification and classification of historical Kannada handwritten document images using LBP features", International Journal of Intelligent Systems Design and Computing, Jan 2018, Vol. 2, Issue 2, pp. 176-188.

[17] T. Ojala, M. Pietikainen, and T. Maenpaa, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," Pattern Analysis and Machine Intelligence, IEEE Transactions on, vol. 24, no. 7, pp. 971–987, 2002.