IJIRCCE

e-ISSN: 2320-9801 | p-ISSN: 2320-9798

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

Volume 9, Issue 7, July 2021

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 7.542

International Journal of Innovative Research in Computer and Communication Engineering

e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 7.542

|| Volume 9, Issue 7, July 2021 ||

| DOI: 10.15680/IJIRCCE.2021.0907206 |

Incorporation of Author-Suggested Corrections in Handwritten Kannada Documents – A Survey

V. B. Pagi[1], Bhargav H K[2], Ganga A Hadagali[3]

Professor, Dept. Computer Science and Engineering, Basaveshwar Engineering College (Autonomous), Bagalkot, Karnataka, India[1]

Asst. Prof. Dept. Information Science and Engineering, Agadi College of Engineering & Technology, Lakshmeshwar, Gadag, Karnataka, India[2]

M. Tech, Dept. Computer Science and Engineering, Basaveshwar Engineering College (Autonomous), Bagalkot, Karnataka, India[3]

ABSTRACT: In the handwritten documents we may come across many corrections like suggesting inclusion in between words, striking out a word, and replacing with an alternative word etc. We need to incorporate the marked corrections in the given document. Sometimes, in a document we may find different kinds of handwriting such as cursive writing and running handwritings. After segmenting the written text that is not part of the running line, if the author is expecting the inclusion of the marked text, we need to move the words appropriately to create space for the text to be inserted. In case the author is expecting to remove striked-out word, but the author is not willing to insert any new word in place of the striked out word, just delete the striked out word and adjust the remaining words and spaces appropriately. If the author is willing to insert a new word in place of the striked-out word, remove the strikeout and insert the suggested word after adjusting the spaces properly. This paper presents the studied literature in this area, emphasizing the correction of Kannada and English documents. Later the work may be extended to other languages, since the basic principles remain the same.

KEYWORDS: handwritten documents, Kannada and English language, segmentation

I. INTRODUCTION

Humans can accurately recognize the handwritten characters if they are neat and clean. It is very easy task for human beings. The same can do easily by the kids also. But the same task for machine is very difficult. Various languages use specific script to write [1].

Handwritten documents comprise a free-form of writing that includes many mistakes. Here, the human error is considered as noise. These noises can be in the form of overwriting, strike-out text which may be in the form of a single line, double line, wavy line, cross lines, etc. Complexity arises when it comes to handwritten text recognition because it is hard for the machine to understand the handwriting of various people as no two persons can have the same handwriting.

Also, the classification of an image to be normal/clear or damaged/striked by a human being is easy when compared to a machine. That is where Handwritten Character Recognition (HCR) comes into picture which uses Optical Character Recognition (OCR) to recognize the handwritten characters. However, Optical Character Recognition (OCR) and text analysis is still under research from past decades.

e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 7.542

|| Volume 9, Issue 7, July 2021 ||

| DOI: 10.15680/IJIRCCE.2021.0907206 |

Image classification using machine learning classifiers will help in reducing the gap between computer vision and human vision. In our context Image classification involves the process of classifying an image into clean/noise-free images and damaged/striked images. Although we see a lot of progress in English OCR, in Indian languages like Kannada no sufficient work is done. Recognition of characters would be challenging as it has curves among its characters.

The fully functional OCR for Kannada handwritten text is still an ongoing process as Kannada characters have many curves which adds to the difficulty in the recognition of the characters but performance decreases with noise in the image/document, as noise cannot be ignored but interpreted as junk. Here we consider noise as struck-out words. As other noises in the images can be handled by cleaning the image using different image processing techniques. This need in the process of improving the OCR detection performance gave us the idea for our project. The results will help in many other applications like writer identification, online evaluation and forensic applications and many more.

II. MOTIVATION

The authors submit their handwritten copies for publishing, suggesting corrections at many places. In the form of strikeouts, insertions and alternative words. Automatic conversion tools fail to incorporate the corrections. Hence, the motivation.

III. OBJECTIVES

The objectives are listed below:

1. To segment strikeouts and corrections and incorporating in the running text.
2. Inclusion of the marked text: move the words appropriately to create room for the text to be inserted.
3. Removal of strike out word, without willing to insert any new word in place of the strike out word
4. After deleting the striked-out word and adjust the remaining words and spaces appropriately.
5. To insert a new word in place of the strike out word, remove the strike out word and insert the suggested word after adjusting the spaces properly.

IV. PROBLEM STATEMENT

To incorporate the corrections suggested by the author for incorporating the strike out words. The corrections may be suggested to words, lines or entire paragraphs.

V. LITERATURE SURVEY

Literature survey / review is also referred to as a systematic review. A form of secondary study that uses a well-defined methodology to identify, analyze and interpret all available evidence related to a specific research question in a way that is unbiased and (to a degree) repeatable.

The motivation behind adopting the literature review is to gain knowledge towards the data sets and the implementation of different types of classifiers to recognize the handwritten characters from the document. Once the required data has been obtained from the literature review, then data analysis is performed.

[1] James A. Thom used handwritten English texts database IAM and modified it to get Struck-out words in training and testing 1900 & 480 and Non-struck-out words in training and testing 14040 & 17080 respectively. Different types of strokes such as single line, double-line, and single diagonal and cross mark are considered. The value of stroke is known from the histogram of the grayscale image. Stroke width can be calculated using Euclidean distance

e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 7.542

|| Volume 9, Issue 7, July 2021 ||

| DOI: 10.15680/IJIRCCE.2021.0907206 |

transform to find the center of the cross strokes and double the average width to find the cross stroke width. 1D bidirectional-LSTM is used to perform the classification. The network consists of five convolution blocks of each 2D convolution layer with the kernel having 3x3 pixel and stride 1x1. (Width x (height x depth))- Column wise concatenation is performed after the final layer. Bidirectional-LSTMs is equal to 80 times the height. Recurrent blocks consist of bidirectional 1D LSTM layers .All five of these layers have 256 units fully connected layers with nodes(number of characters in dataset+1) that are used on the output of the final convolution block. Character Error Rate (CER) on training and Validation is found to be 0.02 and 0.08.Test on the IAM test set achieved 0.09 CER and WER 0.24 and test on the Modified-IAM test struck out text recognition accuracy to be 0.11 CER and 0.25 WER.

[2] Tejashwini Gadag et. al. used handwritten character recognition. In handwritten Character Recognition (HCR) the challenging field is image processing in which there active research is going on. Kannada language is an ancient language which consists of lot of handwritten documents and digitizing them is very helpful to preserve them over time. They have used unconstrained features and CNN, HMM,SVM as classifiers. The dataset is created by their own of 74k characters. After evaluating 92% of accuracy is obtained.

[3] Ramesh. G et. al. in their work propose to work on Handwritten character recognition which is an important subfield of Computer Vision which has the potential to bridge the gap between humans and machines. Machine learning and Deep learning approaches to the problem have yielded acceptable results throughout, yet there is still room for improvement. off-line Kannada handwritten character recognition is another problem statement in which many authors have shown interest, but the obtained results being acceptable. The initial efforts have used Gabor wavelets and moments functions for the characters. With the introduction of Machine Learning, SVMs and feature vectors have been tried to obtain acceptable accuracies. Deep Belief Networks, ANNs have also been used claiming a considerable increase in results. Further advanced techniques such as CNN have been reported to be used to recognize Kannada numerals only. In this work, we budge towards solving the problem statement with Capsule Networks which is now the state of the art technology in the field of Computer Vision. We also carefully consider the drawbacks of CNN and its impact on the problem statement, which are solved with the usage of Capsule Networks. Excellent results have been obtained in terms of accuracies. We take a step further to evaluate the technique in terms of specificity, precision and f1-score. The approach has performed extremely well in terms of these measures also.

[4] P.V. RAMANA MURTHY et. al propose to work on machine learning and neural networks which are trending technologies used in different kinds of handwritten (HW) pattern recognition of various research areas. Therefore it is very hard to distinguish and recognize the hand written characters of different person. Recognition of characters related to telugu language is a part of pattern reorganization that happen to the idea of research during the past some years. Neural networks (NN) are playing a significant role in telugu HW character recognition. HW detection is the capability of a digital computer to receive and understand intelligible HW input from documents, images, touch screens and other electronic devices etc. These all may be online or offline, in this context online recognition includes conversion of pen tip digital movements into a list of originates used as input for the categorization system where as offline recognition uses images of characters like input. For HW reorganization NN has been achieved and improve the efficiency up to 98.3% this is good achievement.

[5] Shakunthala B S et. al. propose to work on Extraction of lines and words from handwritten document images containing skewed text is one of the most difficult and challenging problem. In this paper, a new deskewing algorithm leading to line and word segmentation from an unconstrained hand written Kannada documents is proposed. The method employs preprocessing, dilation and labelling the connected components of input image as initial step. Then an intelligent technique is used to group the words belonging to a text line. The extracted words are subjected to removal of unwanted information that pertains to adjacent words. Further the skew angle computation and rotation operation (when angle is other than zero) are performed for purpose of deskewing of extracted words. Then deskewed words are intelligently written into new image without overlapping of words in text line. The method also takes care of detecting text lines containing consonant modifiers. Inter word and intra character gap variations are also taken care at the time of word segmentation by the proposed method.

e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 7.542

|| Volume 9, Issue 7, July 2021 ||

| DOI: 10.15680/IJIRCCE.2021.0907206 |

[6] Mayur M Patil et. al in their work propose to work on Optical Character Recognition (OCR) which is automatic reading of optically sensed document text materials to translate human-readable characters to machine- readable codes. In Optical Character Recognition, the text lines in a document must be segmented properly before recognition. English Character Recognition (CR) has been extensively studied in the last half century and progressed to a level, sufficient to produce technology driven applications. But same is not the case for Indian languages which are complicated in terms of structure and computations. This is the motivation behind choosing OCR for Kannada language. A KSRTC bus pass application form written in Kannada is chosen for processing and recognition. The OCR system is devised to first segment the whole document into text lines, then to words and then to individual characters. These characters are then used to extract the necessary features and recognize those characters and classify them.

[7] Kiran Y. C et. al. proposed to work on recognition of handwritten text that has been one of the active and challenging areas of research in the field of image processing and pattern recognition. Recognition of Kannada handwritten character is complicated compared to other languages. It has numerous applications which include postal mail application, reading aid for blind and conversion of any handwritten document into electronic form. There is no most robust dataset available for handwritten characters. This paper focuses on developing a dataset for offline handwritten Kannada character recognition and overview of the ongoing researches in this field.

[8] Saleem Pash et. al. in their work propose to work on frontier area of research in the field of pattern recognition and image processing is handwritten character recognition. That leads to a great demand for OCR system containing handwritten documents. In order to recognize the text present in a document, an Optical Character Recognition (OCR) system is developed. In this paper, OCR system for handwritten Kannada characters and numerals is developed which involves several phases such as preprocessing, feature extraction and classification. Preprocessing includes the techniques that are suitable to convert the input image into an acceptable form for feature extraction. The main aim of this paper is to propose an efficient feature extraction and classification techniques. Suitable features are extracted as structural features and wavelet transform is employed for extracting global features. Artificial neural network classifier is used for recognizing the handwritten Kannada characters and numerals. The proposed method is experimented on 4800 images of handwritten Kannada characters and obtained an average accuracy of 91.00%. Also, the proposed method is experimented on 1000 images of handwritten Kannada numerals and obtained an average accuracy of 97.60%.

[9] M. Mahadeva Prasad et. al. in their work is k carried out on recognizing online handwritten data using OLDA. A comparative study on the performance of OLDA and RLDA in terms of recognition accuracy and recognition speed is carried out. Online handwritten Kannada basic character data is used for the experiments. Writer independent experiments are carried with 3750 samples for training and 1550 samples for testing. With estimate feature and nearest neighbor as a classifier, an average maximum recognition accuracy of 88.7% and 88.5% has been achieved with OLDA and RLDA respectively. While OLDA has achieved the best recognition accuracy with only 20 eigen vectors, RLDA has taken 25 eigen vectors. The experiments reveal that the performance of OLDA is better than that of RLDA in terms of recognition accuracy, computation cost and also the memory requirements.

[10] Thungamani.M et. al propose to work on recognition which is well-known fact that building a character recognition system is one of the hottest areas of research as it is shown over the Internet and due to its wide range of prospects. The objective of this paper is to describe an OCR system for handwritten text documents in Kannada. The input to the system is a scanned image of a text and the output is a machine editable file compatible with most typesetting Kannada software. The system first extracts characters from the document image and a set of features are extracted from the character image using Zernike moments. The final recognition is achieved using support vector machine (SVM). The recognition is independent of the size of the handwritten text and the system is seen to deliver reasonable performance.

[11] Sangame S.K et. al. in their work presents unconstrained handwritten Kannada vowels recognition based upon invariant moments. The proposed system extracts Invariant moments feature from zoned images. A Euclidian distance criterion and K-NN classifier is used to classify the handwritten Kannada vowels. A total 1625 image are

e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 7.542

|| Volume 9, Issue 7, July 2021 ||

| DOI: 10.15680/IJIRCCE.2021.0907206 |

considered for experimentation and overall accuracy found to be 85.53%. The novelty of the proposed method is independent of size, slant, orientation, and translation in handwritten characters. Keywords- OCR, Indian Language, Kannada Vowels, Moment invariants

[12] Axel Brink et. al. the method used for automatic identification and removing of crossed-out text in off-line handwriting. It classifies connected components by simply comparing two scalar features with thresholds. The performance is quantified based on manually labeled connected components of 250 pages of a forensic dataset. 47% of connected components consisting of crossed-out text can be removed automatically while 99% of the normal text components are preserved. The influence of automatically removing crossed-out text on writer verification and identification is also quantified. This influence is not significant.

[13] Alessandro Vinciarelli et. al. in their work presents the application of HMM adaptation techniques to the problem of Off-Line Cursive Script Recognition. Rather than training a new model for each writer, one first creates a unique model with a mixed database and then adapts it for each different writer using his own small dataset. Experiments on a publicly available benchmark database show that an adapted system has an accuracy higher than 80% even when less than 30 word samples are used during adaptation, while a system trained using the data of the single writer only needs at least 200 words in order to achieve the same performance as the adapted models.

Table 1. Literature Survey

Year	Authors / Title	Features	Classifier	Datasets	Accuracy
2020	James A. Thom "A Survey on Recognition of Strike-Out Texts in Handwritten Documents"	Hand crafted feature	1D bidirectional -LSTM	NFI dataset	Accuracy is 98.94 %
2020	Tejashwini Gadag et. al. "A Survey on Handwritten Characters Recognition Systems in Kannada Language"	Unconstrai ned	CNN, Hidden Markov Model (HMM)	char74K dataset, SVM	Accuracy is 92 %
2019	Ramesh. G et. al. "Recognition of Off-line Kannada Handwritten Characters by Deep Learning using Capsule Network"	Scalars, Vectors, HOG descriptors	SVM, CNN,	Simulation Data Set	accuracy 94.35%.
2017	Shakunthala B S et. al. "Unconstrained Handwritten Kannada Documents leading	Unconstrai ned, vertical projection profile and structural	Nearest neighborhoo d	Own data set is created of 166 lines and 823 words.	line segmentation accuracy of 96.38% and word segmentation

e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 7.542

|| Volume 9, Issue 7, July 2021 ||

| DOI: 10.15680/IJIRCCE.2021.0907206 |

	to Line and Word segmentation"	features, water reservoir, structural and topological features			accuracy of 92.10%
2016	Mayur M Patil et. al "Handwritten Kannada Document Image Processing using Optical Character Recognition"	SURF based features, generic	KNN	Own data set is created	accuracy of any OCR heavily depends upon segmentation phase
2015	Kiran Y. C et. al. "A Comprehensive Survey on Kannada Handwritten Character Recognition and Dataset Preparation	Direction of the stroke, density of the stroke and number of clicks for the character	KNN	Our own database is created for generating the data samples	Accuracy is of 94.4%
2015	Saleem Pasha et. al. "Handwritten Kannada Character Recognition using Wavelet Transform and Structural Features"	Wavelet transform, Structural features	Artificial Neural Network	Dataset:1885 Training: 435 Testing: 1450	Dataset:1885 Training: 435 Testing: 1450
2011	Thungamani. M et. al. "Off-line Handwritten Kannada Text Recognition using Support Vector Machine using Zernike Moments"	Zernike moments used as a feature vector	Support Vector Machines (SVM), Minimum Mean Distance (MMD), and Nearest Neighbor (NN)	Dataset: 7,410 samples	Accuracy is 94 %

e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 7.542

|| Volume 9, Issue 7, July 2021 ||

| DOI: 10.15680/IJIRCCE.2021.0907206 |

2010	Leena Ragha et. al. "Adapting Moments for Handwritten Kannada Kagunita Recognition"	Unconstrained and isolated OCM, O-GMCM, G-CM and G-GMCM and G-CM+G-S	MLP and BP	Data set: 510 Kagunita Kannada characters from 15 samples.	Accuracy: 69% for vowels and 43% for consonants.
2009	Sangame S.K et. al. "Recognition of Isolated Handwritten Kannada Vowels"	Centric image, Selected moments or other shape measurements	K-Nearest-Neighbor (KNN)	Data set: 1625 1300 samples for training and 325 samples for testing purposes	Accuracy is 85.53%
2008	Axel Brink et. al. "Automatic removal of crossed-out handwritten text and the effect on writer verification and identification"	Branching features and size features	Decision tree, linear support vector machine and k – nearest neighbour	NFI Dataset it consists of 3500 handwritten samples. Train set: 250 Test set: 3250	Accuracy 80.3 Can remove 47% of crossed text 99% of the normal text is preserved

VI. METHODOLOGY

As handwritten English and Kannada text, kannada characters are considered as the challenge due to various factors like enormously large character set, variation in writing style and mood of each individual, size of characters, quality of pen, aging of documents, quality of paper, color of ink, etc. Hence, it is suitable for preprocessing techniques, segmentation, novel feature extraction technique which is proposed for the recognition of handwritten Kannada characters.

e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 7.542

|| Volume 9, Issue 7, July 2021 ||

| DOI: 10.15680/IJIRCCE.2021.0907206 |

VII. PROPOSED WORK

a. BLOCK DIAGRAM

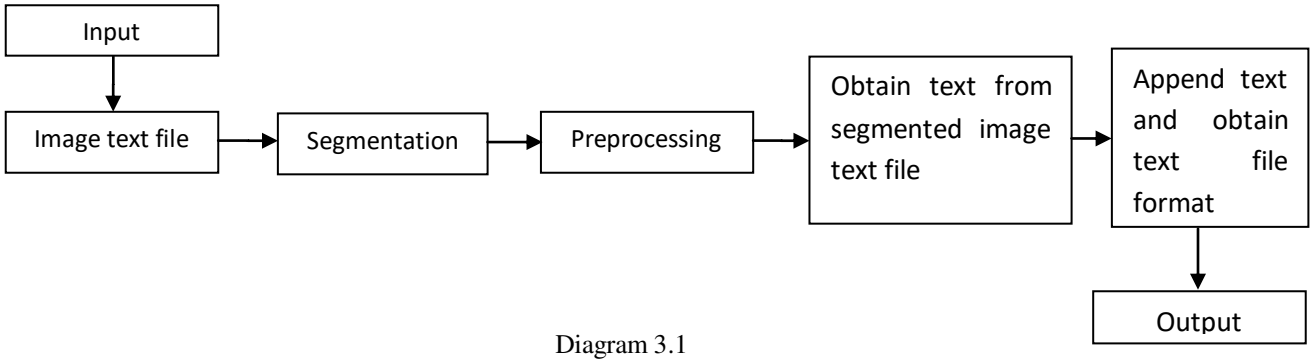

Diagram 3.1

b. FLOW OF WORK

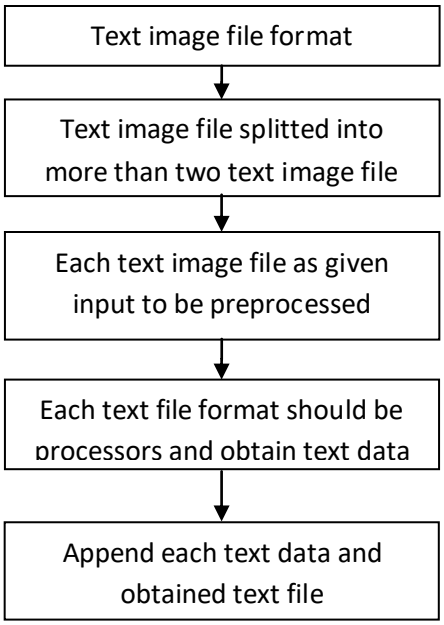

Diagram 3.2

VIII. CHALLENGES

Several challenges are identified and those are listed below:

1. Segmentation is a big challenge due to running handwriting
2. Style of the pen or ink diffusion
3. Subscripts and superscripts for the written text in languages like Kannada and Telugu
4. Joint letter conjugates
5. Strike-out text also appears like the main text

e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 7.542

|| Volume 9, Issue 7, July 2021 ||

| DOI: 10.15680/IJIRCCE.2021.0907206 |

6. Corrections do not appear on the same running line

IX. CONCLUSION

In this paper, we reviewed a number research articles available in literature which discusses the incorporating of author suggested corrections in the handwritten documents in Kannada language. Each of the system use different dataset and different classifier model. We notice that none of the reviewed works has able to achieve very high accuracy. We conclude that in Kannada language needs robust systems which are able to classify the characters with very high classification accuracy.

REFERENCES

[1] Chandana S Upadya, Harshitha H Prabhu, Isiri B N , Dheemanth Urs R A Survey on Recognition of Strike-Out Texts in Handwritten Documents International Research Journal of Engineering and Technology (IRJET) Volume: 07 Issue: 03 | Mar 2020 PP 3690 – 3693

[2] Tejashwini Gadag ,Veena G.S , Jayalakshmi D. S A Survey on Handwritten Characters Recognition Systems in Kannada Language International Journal of Advanced Science and Technology Vol. 29, No. 03, (2020), pp. 12109 – 12114

[3] Hiqmat Nisa, James A. Thom, Vic Ciesielski, Ruwan Tennakoon A deep learning approach to handwritten text recognition in the presence of struck-out text

[4] Madhuri Maheshwari, Deepesh Namdev, Saurabh Maheshwari A Systematic Review of Automation in Handwritten Character Recognition International Journal of Applied Engineering Research ISSN 0973-4562 Volume 13, Number 10 (2018) PP 8090-8099

[5] A.Sakila , Dr. S.Vijayarani Skew Detection and Correction in the Document Image International Journal of Innovative Research in Science, Engineering and Technology Vol. 6, Issue 8, August 2017 PP 17457 – 17465

[6] Shahnaz Abubakker Bapputty Hajia , Ajay James , Dr. Saravanan Chandran A Novel Segmentation and Skew Correction Approach for Handwritten Malayalam Documents International Conference on Emerging Trends in Engineering, Science and Technology (ICETEST - 2015) Procedia Technology 24 (2016) PP 1341 – 1348

[7] RAMANDEEP KAUR, SEEMA BAGHLA, SUNIL KUMAR A REVIEW ON SKEW DETECTION AND CORRECTION IN MULTISCRIPT TEXT DOCUMENT IMAGES International Journal of Advances in Science Engineering and Technology, ISSN: 2321-9009 Volume- 3, Issue-3, July-2015 PP 145 – 148

[8] Saleem Pasha, M.C.Padma Handwritten Kannada Character Recognition using Wavelet Transform and Structural Features International Conference on Emerging Research in Electronics, Computer Science and Technology – 2015 PP 346 – 351

[9] Neha.N LANGUAGE INDEPENDENT ROBUST SKEW DETECTION AND CORRECTION TECHNIQUE FOR DOCUMENT IMAGES International Journal of Electronics Signals and Systems (IJESS) ISSN: 2231-5969, Vol-2, Iss-2 PP 111 – 115

[10] Chandranath Adak, Bidyut B. Chaudhuri An Approach of Strike-through Text Identification from Handwritten Documents 2014 14th International Conference on Frontiers in Handwriting Recognition PP 643 – 648

[11] Axel Brink Harro van der Klauw Lambert Schomaker Automatic removal of crossed-out handwritten text and the effect on writer verification and identification Conference Paper in Proceedings of SPIE - The International Society for Optical Engineering · January 2008 PP 1 – 12

[12] L. Likforman-Sulem, A. Vinciarelli HMM-based Offline Recognition of Handwritten Words Crossed out with Different Kinds of Strokes January 2008 PP 1 – 5

[13] Sangame S.K., Ramteke R. J. , Rajkumar Benne Recognition of isolated handwritten Kannada vowels Advances in Computational Research, ISSN: 0975–3273, Volume 1, Issue 2, 2009, pp-52-55

[14] Kiran Y. C, Lothitha B. J A Comprehensive Survey on Kannada Handwritten Character Recognition and Dataset Preparation International Journal of Engineering Research & Technology (IJERT) ISSN: 2278-0181 Published by, www.ijert.org ICESMART-2015 Conference Proceedings Special Issue – 2015 Volume 3, Issue 19 pp 1 – 4

INNO SPACE
SJIF Scientific Journal Impact Factor
Impact Factor: 7.542

doi cross ref

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

9940 572 462 6381 907 438 ijircce@gmail.com

www.ijircce.com

Scan to save the contact details