International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 07 Issue: 03 | Mar 2020 www.irjet.net p-ISSN: 2395-0072

# A Survey on Recognition of Strike-Out Texts in Handwritten Documents

**Chandana S Upadya[1], Harshitha H Prabhu[2], Isiri B N[3] , Dheemanth Urs R[4]**

*[1,2,3]Student, Dept. of Information Science and Engineering, Dayananda Sagar College of Engineering, Bangalore, Karnataka, India*
*[4]Assistant Professor, Dept. of Information Science and Engineering, Dayananda Sagar College of Engineering, Bangalore, Karnataka, India*

---

**Abstract –** A lot of research work is done in the field of OCR during the last few decades. Complexity arises when it is handwritten and increases if noise exists as it is a barrier for the recognition. In this survey, as to resolve this issue in Kannada we investigate the various work done in the field of Recognition of strike-out text and removal of strike done in various other languages like English, Bengali, Devanagari etc. This survey is of 3 sections. Section 1 is about introduction. Section 2 has a brief explanation of all the research work done in this field and we compare the results of their research. In section 3, we conclude the survey.

***Keywords***: **Document Analysis, Handwritten documents, Optical Character recognition, Strike out texts, Information Retrieval.**

## 1. INTRODUCTION

Handwritten documents comprises a free-form of writing that includes many mistakes. Here, we call this human error 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.

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 which can be seen in Fig 1.

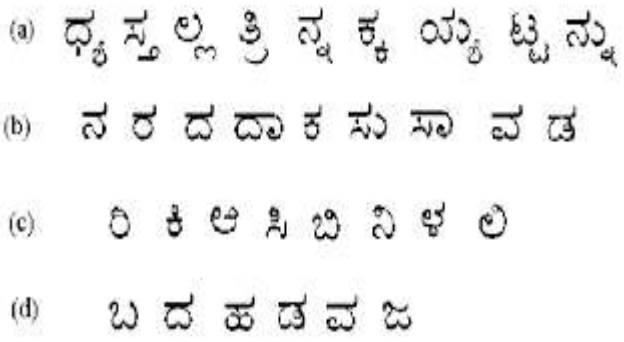


**Fig.1**. Kannada characters with curves.

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.

There are already few works done in this field in languages like Bengali, English, Devanagari etc. There is a need for such experiments in every language as long as people use it for exchange of information. Kannada is a Dravidian language spoken not only by the people of Karnataka but also to some extent by the neighbouring states of Karnataka. The Kannada literature dates from the ninth century and has 47 characters in its alphabet set, 13 vowels and 34 consonants. A word is created by the combination of these vowels and consonants which are called aksharas. The main drawback is the dataset, the insufficient or unavailability of the standard datasets. Even though some kannada printed data, numeric and character datasets are available, handwritten kannada dataset is still scarce. This is causing a lot of researchers to step back from exploring the possibilities.

The detection and removal of the strikes or the struck-out words will help in understanding the handwritten

documents in a better way. Noises such as ink spills, torn pages, over writings can be seen in handwritten scripts.

Considering strike-out texts as noise, we will be briefing out the important details from different experiments done by various researchers on languages like English, Bengali, Devanagari etc. These results motivated us to research our native language Kannada.

## 2. LITERATURE SURVEY

James A. Thom [1] 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 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 convolutional 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. A 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. Three models were used with two training from scratch and one with model1.Two models Modified-IAM Dataset and one with IAM dataset. Drawbacks are model1 lowered the performance using modified IAM and the dataset was insufficient. Model 2 was overfitted. Model 3 identified almost 439 struck out words out of 480 words.

Laurence Likforman-Sulem[2] proposed a method to recognize the crossed out handwritten words. This crossed out word or strike-through word could be done by a single line, double line, wavy line, etc. There are various kinds of strokes. Here, they considered mainly two types of strikes, i.e., wavy trajectory strokes and line trajectory strokes.

The recognition of handwritten words were done using a model known as the Hidden Markov Model. The simulated strokes were superimposed to the original clean word images of the documents to get a clear view of images without strikes and noise-free. The simulation was done using two approaches. Firstly, for wavy trajectory strokes, it is superimposed to clean word images which were created with control points and spline curves. Secondly, for line trajectory strikes, it is superimposed to clean word images of horizontal lines which was generated with the delta-lognormal model of rapid line movements. The recognition approach is by binarization followed by the normalization of images of the words. This normalized word is converted into a sequence of feature vectors, $X=(x_1,\ldots\ldots x_N)$ called as the sliding window approach. From this sequence X, the recognition process can be considered as finding the word model that maximizes the likelihood. This estimation is done using a continuous density Hidden Markov model.

The Hidden Markov Model is trained using the Baum-Welch algorithm that implements the Expectation algorithm. The model is trained with two kinds of datasets called as the strike-free words and tested and validated using the noisy and strike-free words and is validated using the testing set. While tested, normal text gave 91.9% recognition rate while single striked (L1) text gave recognition rate 79.9% and the wave (W) striked text gave recognition rate 45.7%.

B. B. Chaudhary [3] had proposed a model which included tasks like, Identification and localization of Strike-out Strokes and Cleaning by removing the strokes.The datasets are of uncontrolled data, controlled data and semi controlled data. Controlled data included strokes like Single, Multiple, Slanted, Crossed, Zig-zag and Wavy. Mainly four steps are performed pre-processing the image, strike-out word detection using SVM followed by strike-out stroke detection by graph path finding and cleaning of strike-out words by image inpainting which can be seen in the Fig-2.

Each word is subjected to a SVM with RBF kernel based 2-class classifier and the 2 Classes used for this classifier are non-struck-out (class-1) words and struck-out. Inpainting can be performed with the help of mask region. The morphologically dilated version of strokes is considered a mask. Finally binarizing the image post inpainting.1432 Bengali struck out words are used and the accuracy obtained was 94.77%. (class-2) words.

The accuracy results of different classification models developed by B B Chaudhary in classifying the strike out from normal texts in Bengali language can be viewed in Table-1.

Axel Brink[4] proposed a model which was trained and tested on NFI, a forensic dataset. It has 3500 handwritten documents taken from criminal suspects. The handwriting here has many striked-out texts from criminals. This model has 3 stages of training and testing. First stage consists of the 250 pages of NFI dataset where classification is assessed based on the number of

connected components in it. Second stage included assessing the writer verification and the third stage included the writer identification which together used 2374 pages. The initial 250 pages from the NFI dataset was applied to Otsu's thresholding method. And the connected components were fetched using the concept of 8-connectivity. This resulted in fetching 86537 connected components. Based on the connected components the dataset was classified into 3 different groups: normal, crossed and other. The 'other' group had text with connected components that could not be figured out as normal text or is a noise.

Decision tree was used to classify the striked out words based on the features of connected components and by setting threshold. The two features that helps separate them are branching feature and size feature. The text with threshold both above the size feature and branching feature is considered to belong to the striked out word. This involved using the first 1-125 page for training and 126-250 page set for testing. The resulting images were normalized and a 2 dimensional data was obtained that was labelled. The threshold of branching feature (θb) = 1.5 and size feature (θs) = 1 was used to classify the crossed out words from normal and other words. The result found had True Positive=47.5% and True negative = 99.1%. The resulting documents were used for writer identification and verification. Based on hinge feature and $\chi$ 2 -distance, writer verification was performed. The writer identification was done by obtaining a hit list based on the matching and grouping the pages with closer hinge feature and $\chi$ 2 -distance.

B. B. Chaudhary [5] found that the strike-out on a text forms the connected component. So, they have labelled the connected components from the binarized text.They have also neglected the connected components that have black pixel's count less than the threshold of noise (Tn).The start and end of the strokes are found better by thinning the image and representing it in the graph format. G = (V, E), where V is its set of nodes and E is its set of edges. Where nodes and edges are located by traversing in horizontal/vertical direction and diagonal direction.

As these strokes are considered as the connected components, they considered the ends of the strokes as the terminal nodes namely, left and right nodes with three regions of the word as left, middle and right regions. This stroke could be the shortest edge between the left and right node. So, they considered a graph formed by discarding self-loop and having the graph with less shortest paths. If VL1 and VR1 are left and right nodes, then VL1 will not have 8-neighbor pixels on its left and VR1 will not have 8-neighbor pixels on its right. This way, the connected component is found. The shortest path between the nodes was found using Dijkstra's algorithm.

The entries in the distance matrix represent the strokes (Assumed to be a straight line). The Euclidean line and shortest path is drawn between nodes and a perpendicular is drawn between them. It is considered to be a stroke if these perpendicular distances are less than a threshold. If no path is found then there are no strokes found. There is a need for a special method to handle the lines which are part of the character/text from being misinterpreted as SS. This can be handled by taking the crossing count above and below the matra. Where this count for SS will be nearly the same on both sides. This property distinguishes matra/shirorekha from the strokes. After removing the SS by ignoring the intersection points with the character, thickening of the image is done with morphological dilation operation. Thus returning the original image.130 of the total handwritten documents are created by different people.

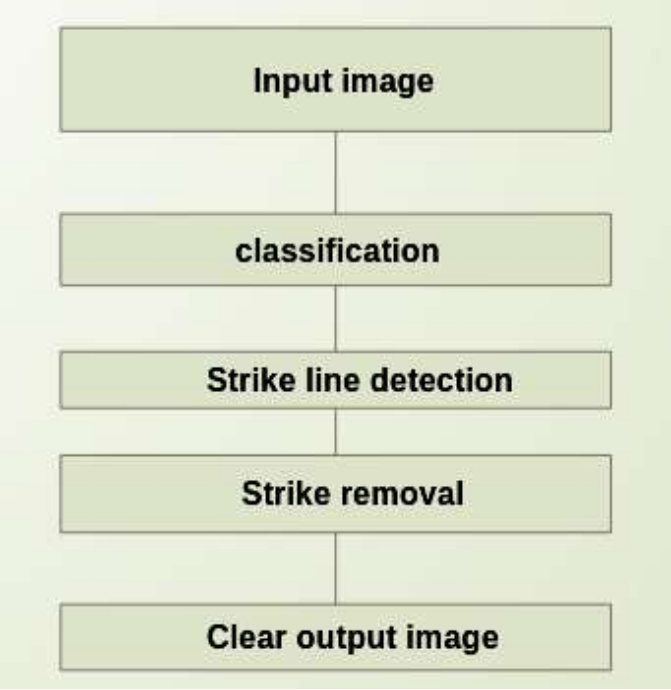


**Fig 2:** Sequence diagram for model proposed by B. B. Chaudhuri

**Table-1:** Comparison of results from different classification models

| Method | Precision % | Recall % | F-Measure % |
|---|---|---|---|
| Hand-crafted feature + SVM | 90.19 | 91.94 | 91.06 |
| CNN feature + SVM | 97.25 | 97.84 | 97.54 |

The Table 2 denotes the classifier techniques and the accuracy of each method implemented by the authors along with the language they worked on.

**Table 2:** Compared results from different research works**.**

| Language | Author | Classifier | Accuracy |
|---|---|---|---|
| Bengali | B.B.Chaudhuri | CNN feature + SVM | 98.2 |
| English | James A. Thom | 1D bidirectional-LSTM | 98.94 |
| English | Axel Brink | Decision tree | 80.3 |
| English | Laurence | HMM | 70.1 |

## 3. CONCLUSION:

In this paper, we discussed various methods suitable for classifying striked text from normal text and cleaning techniques. Results from various methods are compared and tabulated.

## REFERENCES:

1. Nisa, Hiqmat & Thom, James & Ciesielski, Vic & Tennakoon, Ruwan. (2019). A deep learning approach to handwritten text recognition in the presence of struck-out text. 1-6. 10.1109/IVCNZ48456.2019.8961024.
2. Likforman-Sulem, Laurence and Alessandro Vinciarelli. "HMM-based Offline Recognition of Handwritten Words Crossed Out with Different Kinds of Strokes." (2008).
3. Mioulet, B. B. Chaudhary, C. Chatelain, T. Paquet, "Unconstrained Bengali handwriting recognition with recurrent models", Proc. ICDAR, pp. 1056-1060, 2015
4. Brink, Axel & Klauw, Harro & Schomaker, Lambert. (2008). Automatic removal of crossed-out handwritten text and the effect on writer verification and identification. 6815. 68150. 10.1117/12.766466.
5. Adak, Chandranath & Chaudhary, Bidyut. (2014). An Approach of Strike-Through Text Identification from Handwritten Documents. Proceedings of International Conference on Frontiers in Handwriting Recognition, ICFHR. 2014. 10.1109/ICFHR.2014.113.
6. A. Priya, S. Mishra, S. Raj, S. Mandal and S. Datta, "Online and offline character recognition: A survey," 2016 International Conference on Communication and Signal Processing (ICCSP), Melmaruvathur, 2016, pp. 0967-0970.
7. Boiangiu, Costin-Anton & Raducanu, Bogdan. (2008). Robust Line Detection Methods.
8. S. Nicolas, T. Paquet, L. Heutte, "Markov random field models to extract the layout of complex handwritten documents", in: IWFHR, 2006.
9. D. Tuganbaev, D. Deriaguine, "Method of Stricken-out Character Recognition in Handwritten Text", Patent US 8,472,719, 25 June 2013.
10. Kiran Y. C, Lothitha B. J, 2015, A Comprehensive Survey on Kannada Handwritten Character Recognition and Dataset Preparation, INTERNATIONAL JOURNAL OF ENGINEERING RESEARCH & TECHNOLOGY (IJERT) ICESMART – 2015.
11. Thungamani M, Dr.Ramakhanth Kumar, P KeshavaPrasanna, Shravani Krishna Rau Off-line Handwritten Kannada Text Recognition using Support Vector Machine using Zernike Moments, International Journal of Computer Science and Network Security, July 2011.
12. Nilanjana Bhattacharya, Volkmar Frinken, Umapada Pal, Partha Pratim Roy, "Overwriting repetition and crossing-out detection in online handwritten text", Pattern Recognition (ACPR) 2015 3rd IAPR Asian Conference on, pp. 680-684, 2015.
13. Chandranath Adak, Bidyut B. Chaudhary, Michael Blumenstein, "Impact of struck-out text on writer identification", Neural Networks (IJCNN) 2017 International Joint Conference, 2017.
14. Plamondon, Réjean & Guerfali, Wacef. (1998). The generation of handwriting with delta-lognormal synergies. Biological Cybernetics.
15. Morasso, Pietro. (1986). Understanding Cursive Script as a Trajectory Formation Paradigm.
16. Niu, Xiao-Xiao & Suen, Ching. (2012). A novel hybrid CNN-SVM classifier for recognizing handwritten digits. Pattern Recognition. 45. 1318-1325. 10.1016/j.patcog.2011.09.021.
17. Vinciarelli, Alessandro & Bengio, Samy & Bunke, Horst. (2004). Offline Recognition of Unconstrained Handwritten Texts Using HMMs and Statistical Language Models. IEEE transactions on pattern analysis and machine intelligence.