Identifying Strikeouts and Predicting Synonym Words for Kannada Handwritten Documents

Bhargav H K[1]
[1] Director, IIIC and Assoc.Professor, Department of Computer Science and Engineering,
Shridevi Institute of Engineering and Technology, Tumakuru-572 106
bhargavwin@gmail.com

Abstract

Handwritten character recognition (HCR) is a pivotal facet within the realm of machine learning and artificial intelligence. This research delves into the intricate challenges associated with detecting and recognizing Kannada handwritten document images. The paper introduces a novel methodology designed to identify strike-outs, addressing a significant aspect of document analysis. In this research, we employ dilation and contour analysis techniques to efficiently discriminate between sentences and individual strikeout words, enhancing the accuracy of our recognition system. Furthermore, we present an innovative synonym generation mechanism for the identified strikeout words, augmenting the utility of our approach in the context of document understanding and information retrieval. This work contributes to the advancement of HCR, especially in the context of Kannada script, and underscores the significance of tackling nuanced aspects such as strike-outs for comprehensive document analysis and interpretation.

Keywords: Handwritten character recognition, machine learning, Dilation, Contour, strike-outs, Identification

1. Introduction

Computer Vision is rapidly evolving due to ongoing advances in machine learning and deep learning techniques. Within the realm of natural language processing, the development of efficient and versatile handwritten text and document recognition poses a formidable challenge. This domain finds applications in linguistic script analysis, detection, recognition, translation, signboard interpretation, archaeological research, and aiding visually impaired individuals. Additionally, it plays a pivotal role in digitizing extensive document archives, facilitating seamless access to historical records, and preserving valuable information [1].

Handwriting recognition systems are typically categorized into two primary modes: online and offline. Optical Character Recognition (OCR) serves as a prominent example of the offline mode. Recent research endeavors have focused on recognizing words, phrases, and sentences from high-resolution digital images of handwritten documents in various languages. In contrast, the online mode involves capturing and analyzing the dynamic pen tip movements as characters are handwritten. This study primarily concentrates on Kannada handwritten text and document images.

Kannada, a regional language of the Karnataka state in India, holds the esteemed status of being an ancient classical language recognized by the Government of India. It boasts a rich heritage with a wealth of ancient handwritten scripts in need of digitization. Kannada comprises 49 distinct characters, including 13 vowels, 34 consonants, and yogavahakas. Recognizing handwritten Kannada text using advanced machine learning or deep learning techniques is particularly intriguing and challenging due to the significant variability in individual handwriting styles and the absence of consistent spacing between letters, words, and lines

One of the most pressing research challenges in Kannada handwriting recognition is the scarcity of suitable standard datasets for developing recognition models. The intricate nature of the script makes it exceedingly difficult to manually curate datasets that encompass the diverse combinations of Kannada characters. Notably, there is a noticeable absence of research efforts dedicated to the identification and recognition of strikeout annotations in Kannada handwritten text or document images.

This study aims to bridge critical gaps in the area of Kannada HCR by addressing the ability to detect strikeout annotations in handwritten Kannada text and providing suggested synonyms. The following sections of this paper are organized in the following order: Section 2 includes an extensive review of relevant literature. Section 3 explains the approach of our proposed model, including the strategies used for strikeout detection and

synonym creation. Section 4 presents the findings of our experimental research, which provide insight into the performance and effectiveness of our approach. Finally, Section 5 finishes the report with a conclusion and findings from our research.

2. Related works

Deep neural networks have gotten a lot of interest in machine vision during the last decade, owing to their outstanding recognition accuracy. This literature study reaches into the improvement of Convolutional Recurrent Neural Networks (CRNN) for HCR, focusing on lines including struck-out words. The CRNN algorithm was trained on the IAM line database and then tested on lines with struck-out words, which led to a slight rise in the Error Rate from 0.09 to 0.11. Particularly, in recognizing instances of struck-out text, their approach surpassed existing approaches [2]. Another study described a unique technique for recognizing and analyzing struck-out text in images of unconstrained handwritten documents. This method consists of two independent approaches: (i) pattern and (ii) text recognition using a graph-based technique. The first approach uses a Support Vector Machine (SVM). The second approach involves modeling the text as a graph and using a limited shortest-path technique to identify strike-out. This approach was evaluated on a 500-page dataset and yielded an impressive F-Measure of 91.56% of struck-out detection in English text [3].

The Kannada HCR was used in another work to use deep learning methods. For categorization, Convolutional Neural Networks (CNNs) were employed, which provided a stable and efficient method of recognizing handwritten letters. The model performed admirably, obtaining roughly 93.2% and 78.73% accuracy for two various datasets [4]. Similarly, in the next work for Kannada HCR, the CNN and the Tesseract tool achieve 87% and 86% accuracy [5]. The Kannada language, with its numerous and complex characters, poses a unique difficulty when it comes to character extraction. A dataset with ten separate classes was used in another study for an experimental evaluation. A unique technique using Random Forest and SVM algorithms was developed, resulting in an amazing classification accuracy of 78% [6]. Another investigation used a deep learning CNN algorithm to recognise non-overlapping lines of Kannada characters. This model was trained on the Chars74K database and obtained a 98% accuracy in classification [7].

In addition, an Artificial Neural Network (ANN) was used to recognize Kannada numbers and handwriting characters, along with a wavelet transform for global feature extraction. This method was evaluated on a large dataset of 4800 and 1000 Kannada characters and numerals, giving excellent accuracy in classification of 91% and 97.60% [8]. For recognizing Devanagari (Hindi) handwriting characters, the algorithm used a Histogram of Oriented Gradients (HOG). The workflow of the model involved segmentation, pre-processing, feature extraction, categorization, and identification. An ANN was used to classify individual characters, yielding an impressive categorization accuracy of 97.06% [9]. Additionally, Urdu character recognition used the UNHD dataset and a two-stage procedure, including feature extraction carried out with a CNN algorithm and classification achieved using a bi-directional Long-Short Term Memory technique. This seven-layer multi-layered technique attained a classification accuracy of more than 83% [10].

3. Methodology

The methodology utilized in this work is depicted in the block diagram shown in Figure 1. This approach consists of three distinct steps, all aimed at the detection of strikeouts and the provision of suggested synonyms.

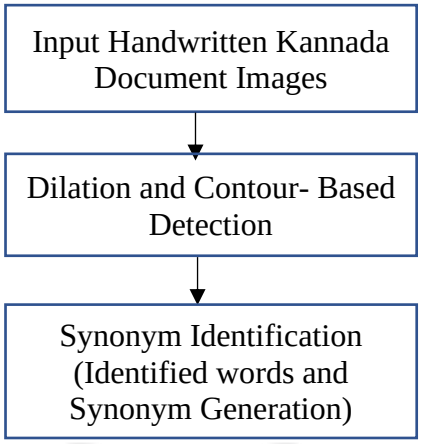

Fig. 1: Flow of proposed methodology

Step 1: Image Preprocessing

Our process initiates with the acquisition of an image containing handwritten text, often characterized by lighting, color, and background variations that can introduce complexity. Therefore, our first step involves image preprocessing techniques to ready the image for more intricate operations. A pivotal aspect of this stage is the conversion of the input image into grayscale. This initial transformation simplifies subsequent analysis by reducing the image's complexity to a single channel representing pixel intensity. Grayscale conversion also mitigates the impact of color variations that may exist in the original image. The significance of this step lies in its role as the foundational stage for all subsequent operations. A properly preprocessed image establishes a clear and consistent starting point for the detection and interpretation of handwritten text, which is the ultimate goal of our methodology.

Step 2: Dilation and Contour-Based Detection

The grayscale image undergoes two critical operations: dilation and contour-based detection. This step is pivotal for precise strikeout detection within the handwritten text. Dilation is applied to the grayscale image, a morphological operation that expands the boundaries of ink strokes in the text, rendering the text characters more distinct from the background. This process results in enhanced contrast between the text and its surroundings, simplifying the identification and delineation of the handwritten content. Following dilation, contour identification techniques come into play. Contours represent the outlines of distinct objects or regions within an image. In our context, they play a vital role in segmenting the text, allowing for the accurate localization of individual sentences and strikeout annotations. Figures 2 and 3 visually illustrate the effectiveness of our approach in identifying sentences and strikeout words through contour-based detections. The synergy between dilation and contour analysis empowers our methodology to precisely pinpoint the regions of interest within the handwritten text.

Step 3: Synonym Identification

The final phase of our methodology entails enriching the analysis through synonym identification. While Step 2 adeptly identifies handwritten words, including those that are struck out, our approach transcends mere identification; it enhances the semantic comprehension of the handwritten content. In this step, the words identified in the previous stage are subjected to a synonym generation process. This process leverages language models and lexical databases to discover alternative word choices for the struck-out text. By providing synonyms, our approach not only acknowledges the presence of strikeout annotations but also offers meaningful insights into potential alternatives for the crossed-out words.

4. Experimental Analysis

Our methodology underwent rigorous experimental analysis on a computer system featuring an Intel i5 processor, 4 GB of RAM, and a 1 TB hard disk, yielding significant and noteworthy results. To ensure the relevance and specificity of our research within the Kannada handwritten text recognition domain, we meticulously curated a customized dataset.

Figures 2, 3, and 4 serve as pivotal visual representations of our experimental findings. Figures 2 and 3 emanate from the application of image dilation and contour techniques, showcasing the remarkable effectiveness of our approach in identifying sentences and individual words within handwritten text. These figures demonstrate how these two strategies work together to accurately isolate text regions and establish distinct boundaries. The accuracy in text localization is critical for later stages of our research, such as strikeout annotation detection and synonym recognition.

Figure 4 shows the outcome of our methodology's strikeout annotation recognition. It is critical to emphasize that our ability to recognize strikeout annotations contributes significantly to the area of handwritten Kannada text recognition. Figure 4 demonstrates the successful isolation of strikeout terms.

Our findings not only validate our methodology's success in recognizing and comprehending handwritten Kannada text, but also show its potential uses in document verification, analysis, and the ability to assist those with visual impairments. Furthermore, the development of our customized dataset was critical in assuring the robustness and accuracy of our findings. By creating our dataset, we tailored our methodology to the unique characteristics of Kannada script, which encompasses 51 characters, including 16 vowels and 35 consonants. This dataset construction also addressed the challenge posed by the variability in individual handwriting styles, a common feature in handwritten documents.

Our rigorous experimental analysis reaffirms the viability and efficacy of our approach in identifying sentences, words, and, notably, strikeout annotations within handwritten Kannada text. Figures 2 and 3 vividly illustrate the accuracy and practicality of our methodology. Our work not only contributes significantly to the field of Kannada handwritten text recognition but also extends its applicability to broader domains such as document analysis and improved accessibility for individuals with visual impairments.

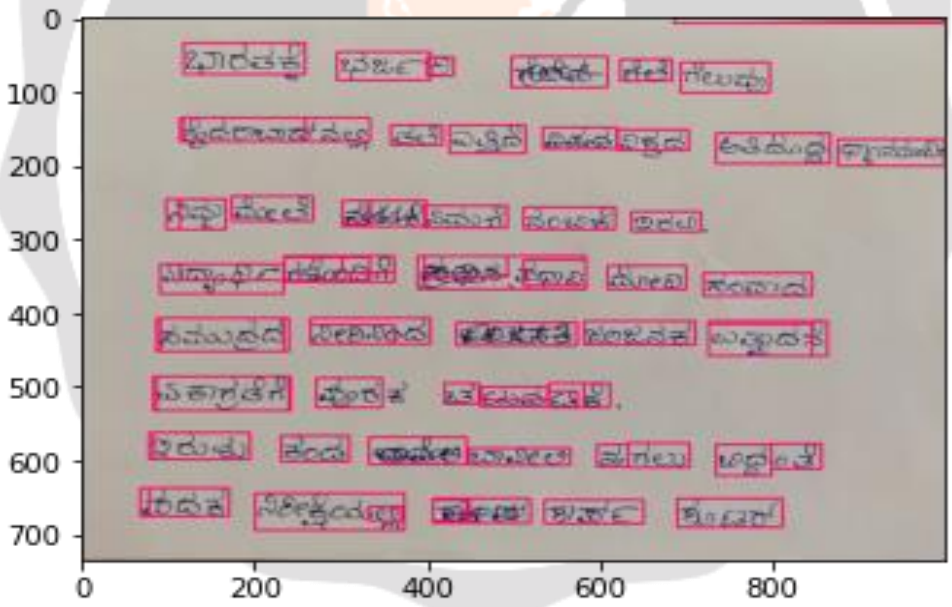

Fig. 2: Words Identification

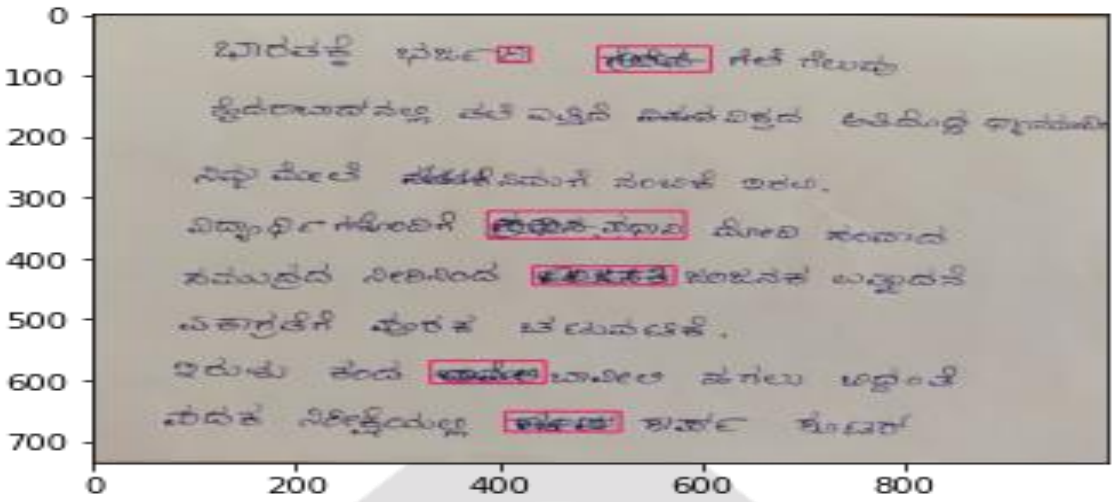

Fig. 3: Strikeout identification

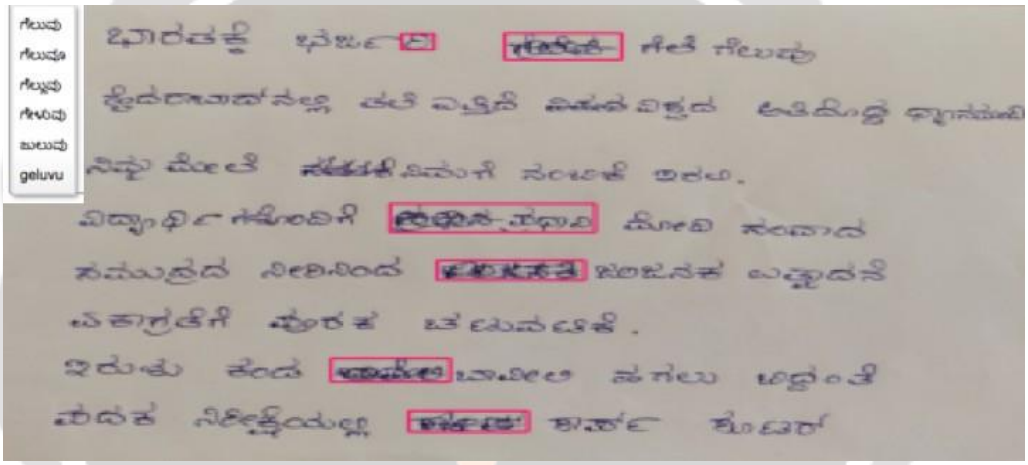

Fig. 4: Synonym Identification

5. Conclusion and Future Work

In summary, this study has introduced a robust methodology for detecting strikeouts in Kannada handwritten documents, complemented by an innovative synonym identification feature. Comprising three essential steps—image preprocessing, dilation and contour-based detection, and synonym identification—our approach has demonstrated its effectiveness in accurately identifying regions of interest within handwritten documents. It sheds light on annotations and provides meaningful insights into potential revisions. Looking ahead, there are several promising avenues for future research and development in this domain. Firstly, extending our methodology to support multiple languages beyond Kannada would enhance its applicability and impact, making it suitable for a broader range of handwritten documents. Secondly, there is an improvement in the quality and accuracy of synonym suggestions. This could be achieved through advanced natural language processing techniques and the incorporation of larger lexical databases, ultimately enhancing the utility of our approach. Lastly, applying our methodology to historical documents could yield valuable insights into language evolution, cultural preservation, and historical research. Our methodology represents a valuable contribution to the analysis of handwritten text, and its ongoing refinement and expansion into new languages and domains hold significant promise for the future of document analysis and information retrieval in handwritten materials.

References

1. Ramesh G, Kumar N. S, Champa HN (2020) Recognition of Kannada Handwritten Words using SVM Classifier with Convolutional Neural Network. In: 2020 IEEE Region 10 Symposium (TENSYMP). pp 1114–1117

2. Nisa H, Thom JA, Ciesielski V, Tennakoon R (2019) A deep learning approach to handwritten text recognition in the presence of struck-out text. In: 2019 International Conference on Image and Vision Computing New Zealand (IVCNZ). pp 1–6
3. Chaudhuri BB, Adak C (2017) An approach for detecting and cleaning of struck-out handwritten text. Pattern Recognition 61:282–294 . https://doi.org/10.1016/j.patcog.2016.07.032
4. Ramesh G, Sharma GN, Balaji JM, Champa HN (2019) Offline Kannada Handwritten Character Recognition Using Convolutional Neural Networks. In: 2019 IEEE International WIE Conference on Electrical and Computer Engineering (WIECON-ECE). pp 1–5
5. Fernandes R, Rodrigues AP (2019) Kannada Handwritten Script Recognition using Machine Learning Techniques. In: 2019 IEEE International Conference on Distributed Computing, VLSI, Electrical Circuits and Robotics (DISCOVER). pp 1–6
6. N SR, Nair B J B, M R A, M L P (2021) Kannada Confusing Character Recognition and Classification Using Random Forest and SVM. In: 2021 3rd International Conference on Signal Processing and Communication (ICPSC). pp 537–541
7. Asha K, Krishnappa HK (2018) Kannada Handwritten Document Recognition using Convolutional Neural Network. In: 2018 3rd International Conference on Computational Systems and Information Technology for Sustainable Solutions (CSITSS). pp 299–301
8. Pasha S, Padma MC (2015) Handwritten Kannada character recognition using wavelet transform and structural features. In: 2015 International Conference on Emerging Research in Electronics, Computer Science and Technology (ICERECT). pp 346–351
9. Singh N (2018) An Efficient Approach for Handwritten Devanagari Character Recognition based on Artificial Neural Network. In: 2018 5th International Conference on Signal Processing and Integrated Networks (SPIN). pp 894–897
10. Hassan S, Irfan A, Mirza A, Siddiqi I (2019) Cursive Handwritten Text Recognition using Bi-Directional LSTMs: A Case Study on Urdu Handwriting. In: 2019 International Conference on Deep Learning and Machine Learning in Emerging Applications (Deep-ML). pp 67–72