Enhancing Machine Learning Methods for Robust Real-Time Text Classification of Bilingual Documents

Santhosha S G[1], Sridhara Acharya P[2], Sampathkumar S[3] & Rechal[4]
[1] *Research Scholar, Dept. of ICIS, Srinivas University, Mangalore & Associate Professor, Dept of MCA, JNN College of Engineering, Shimoga, Karnataka, India.*
[2] *Research Professor, Institute of Computer Science and Information Science, Srinivas University, Mangalore*
[3]*Assistant Professor, Dept of MCA, JNN College of Engineering, Shimoga.*
[4]*PG Students, Dept of MCA, JNN College of Engineering, Shimoga*

Abstract: **The rapid growth of digital data has led to the widespread creation and storage of digital images containing text. The extraction and use of textual information might be advantageous for various kinds of domains. Text detection in natural images is primarily affected by noise, blur, distortions, font variation, alignments, and orientation. Government forms, mark cards, medical records, business receipts, and other increasingly common bilingual documents have a substantial impact about precision of text detection and recognition. The paper, "Enhanced Machine Learning Methods for Robust Real-Time Classification of Bilingual Documents," emphasizes these challenges and suggests a solution that uses image enhancement techniques to improve an image's appearance and quality. Digital documents are bilingual; hence, extracting information from them is challenging since a computer can read and interpret text written in many scripts, including English and Kannada, within the same documentation. It thus extracts text from documents using optical character recognition (OCR)-based problems. Natural language processing (NLP) to classify text, super-resolution methods to synthesize a high-resolution (HR) image from several low-resolution (LR) images and machine learning models EasyOcr to detect and recognize Text.**

***Keywords*— Image Enhancement, Natural Language Processing (NLP), Optical Character Recognition (OCR), Real-Time Text Recognition.**

I. INTRODUCTION

In today's world, an increasing number of bilingual and multilingual digital documents have created large challenges to automated document classification systems. Many of these documents also have reduced quality due to noise, blurriness and a low resolution common in scanned images. Earlier machine learning-based categorization techniques did not be able to predictively classify well and have a non-redundant return on investment in such field, especially in real-time applications, where speed and efficiency are prioritized. This study's main challenge was real-time adaptability, which would require any model of adaptable learning to update itself in real-time with new data as it becomes available.

One of the world's marvels, the human visual system is considered to be replicable. A system needs to know what its purpose is before it can complete any operation. People use their ears, eyes, and brain to process and respond to information. Machines also use word recognition as eye and speech recognition. The study of optimized machine learning frameworks for bilingual document image classification with real-time adaptation faced several obstacles related due to the intricacy of the datasets, variability of languages and scripts, real-time adaptation, class imbalance, and computational resource management. Solving these challenges is vital to creating a precise algorithm that is efficient enough to classify a bilingual document in real time.

Languages with distinct scripts, such as Kannada and English, display a further challenge. A text document has two properties: script and language. Script refers to the alphabet (or character set) for writing documents, while language refers to the way that use the letters to create valid words and sentences that conform to vocabulary and grammar for a particular language. Some scripts can be used to write in multiple languages, so for cultural and historical reasons, for

each script, there exists a set of languages that can be written using that script. Conversely, with a language, it is common to refer to the language as a standard script, which is usually what it is written with. The English, French, Dutch, German languages, etc. for example, are written with the Latin alphabet, whereas Hindi, Kannada, Sanskrit, Marathi, etc. are written using the Devanagari script. Text detection has text detection process involves two steps: "text localization" and "text extraction and enhancement.".

Natural Language Processing (NLP) is a type of data science that is able to learn and understand text data in an intelligent way. NLP enables users to complete tasks such summary text, translate languages, perform sentiment analysis, complete speech recognition, and extract information, among many other text-based problems.

The super-resolution (SR) method reconstructs an image or sequence at a greater resolution from low-resolution (LR) data. In this project, it will review super-resolution approaches and improvements because it has many uses across education, research, and industries. The basic premise of super-resolution is to combine low-resolution (noisy) images of a scene into a high-resolution image. Hence, it tries to reconstruct an original sentinel image at high resolution from previously observed images for sequential sampling at low resolution.

A Gaussian filter is a standard image smoothing technique that uses a Gaussian (bell-shaped) function to blur an image. It works by averaging pixel values in a neighbourhood, giving more weight to core pixels and less to those farther away. This filter helps minimize Gaussian noise and smooth the image without affecting important characteristics, such as edges. It is specifically useful in the preprocessing steps for tasks like OCR, edge detection, and recognizing the object.

II. RELATED WORK

To extract and detect languages in bilingual documents, Shivani Surana, Komal Pathak, Vidhan Shrivastava, Mahesh T R. et al. [1] used Machine Learning Techniques for Text Extraction and Detection from Images: A Review of the Literature. It applies various machine learning techniques, with backpropagation networks, RNN, and BLSTM, in addition to OCR software like Tesseract and EasyOCR, for text recognition and extraction from images. It is a multi-step process that includes image acquisition, preprocessing, segmentation, recognition, and training to handle font, orientation, and background changes. These methods were deployed for larger real-world applications to automate and improve the accuracy of converting printed. For extracting bilingual answer scripts from low-resolution images Santhosh S.G and Sridhara Acharya P et al. [2] used deep learning techniques, CNN. CNN are effectiveness in handling noise, script variability, and partial distortions in complex document images. To classify Kannada and Devanagari/Sanskrit scripts, the study extracts text features from pre-processed document images with Support Vector Machines (SVM). SVM has high binary classification accuracy, the facility toward interacts with a basic graphical user interface, and effectiveness with smaller datasets, as proposed by Shashank Simha B. K., Rahul M., Jyoti R. Munavalli, Prajwal Anand, et al. [3]. M. Liao, Z. Zou, Z. Wan, C. Yao, and X. Bai, et al. [4] enable the application of Adaptive Scale Fusion (ASF) and Differentiable Binarization (DB) to enhance real-time scene text detection through robust multi-scale feature handling and accurate text region extraction. The challenges with complicated backgrounds and computational load, these modules were applied to improve detection accuracy across a range of text sizes and reduce post-processing complexity. In [5], document images were classified and their performance was compared using CNN for visual features and TF-IDF for textual features. CNN displays that deep learning has uses for document image classification, requiring more processing, with a 93% accuracy rate in recognizing visual patterns.

III. MATHEMATICAL MODEL

1. Input Image and Enhancement:

$$I_{LR} \in R^{H*W*C} \text{---------- (1)}$$

equation (1) represents,
H, W = dimensions
C = color channels
$I_{LR} = \text{low} - \text{resolution image}$
Super-resolution

$$I_{HR} = \text{SR}(I_{LR}) \text{---------- (2)}$$

Equation (2) represents,
SR = super-resolution function
$I_{HR} = \text{enhanced high} - \text{resolution image}$

2. Text Extraction using OCR

   T=OCR $(I_{HR}) = \{\omega_1, \omega_2, \dots \dots \omega_n\}$------- (3)

   T = set of recognized words

   ω_i = extracted word

3. Language Identification

   $f(\omega_i) = \begin{cases} 1 \; if \; \omega_i \; is \; Kannada \\ 0 \; if \; \omega_i \; is \; English \end{cases}$

   Partition:

   $W_{kn} = \{\omega_i | f(\omega_i) = 1\}, W_{en} = \{\omega_i | f(\omega_i) = 0\}$

   Total counts:

   N=|T|, $N_{kn} = |W_{kn}|, N_{en} = |W_{en}|$---------- (4)

   equation (4) represents,

   ω_i = individual word

   f(ω_i) = 1 = Kannada, 0 = English

   N=|T|: Total number of words

   $N_{kn} = |W_{kn}|$: Number of Kannada words

   $N_{en} = |W_{en}|$: Number of English words

4. Percentage of Kannada and English words

   $P_{kn} = \left(\frac{N_{kn}}{N}\right) * 100, P_{en} = \left(\frac{N_{en}}{N}\right) * 100$--------- (5)

   equation (5) represents,

   P_{kn}, P_{en}= Language percentage

5. Classification using Machine Learning

   $\hat{y} = \mu(X)$---------- (6)

   equation (6) represents,

   $\hat{y} \in$ {form, receipt, marksheet, medical record}

IV. METHODOLOGY

The proposed approach creates machine learning (ML) for robust text extraction and classification for bilingual documents printed in English and Kannada scripts. It uses the EasyOCR model, which includes a Convolutional Recurrent Neural Network (CRNN), internally linked to a Connectionist Temporal Classification (CTC) decoder, for fast sequence prediction in OCR. Preprocessing operations such as noise removal and resolution enhancement are performed to enhance the quality of the input image before recognition is carried out. The framework is written in Python while the Optical Character Recognition (OCR) is handled by EasyOCR and logical classification by Unicode-based rules. Every extracted word is separated on the basis of its Unicode range in English and in Kannada. The pre-processed texts are tokenized and word counts based on language-specific stop words are calculated. Percentage composition of languages is found using statistical analysis. This systematic process enables script-aware document classification in real-time on Bilingual data.

The workflow diagram displays the steps in taking bilingual text from an image, from the upload of the image to the OCR processing, language extraction and the calculation of percentages. It helps show how EasyOCR and the Flask web interface work together to analyse documents.

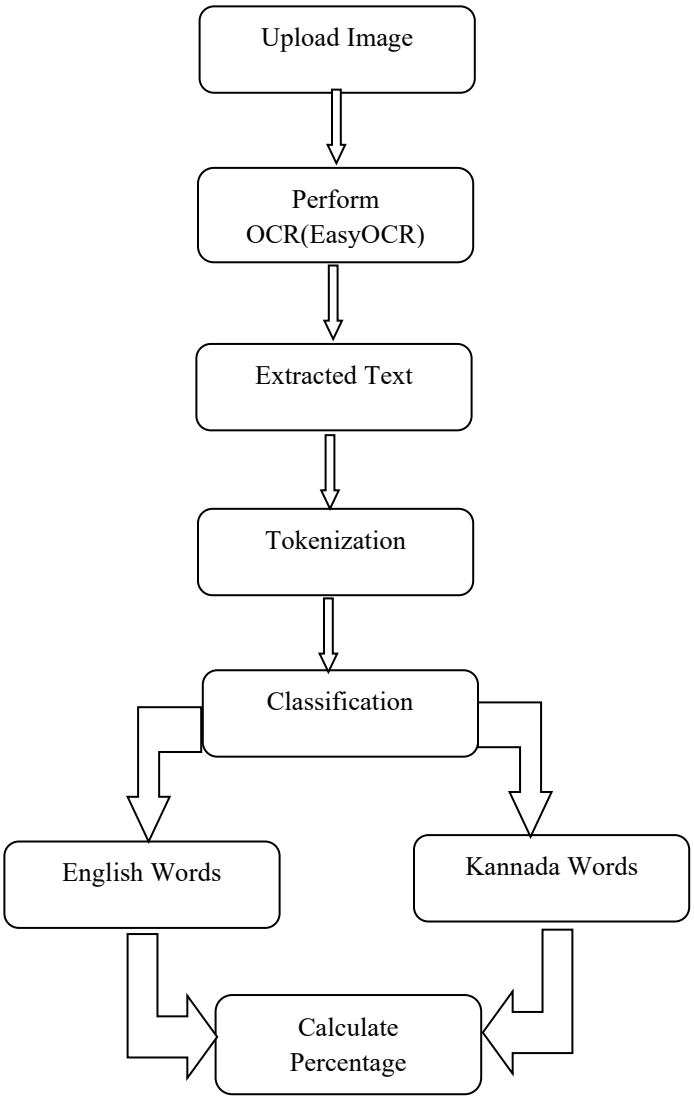
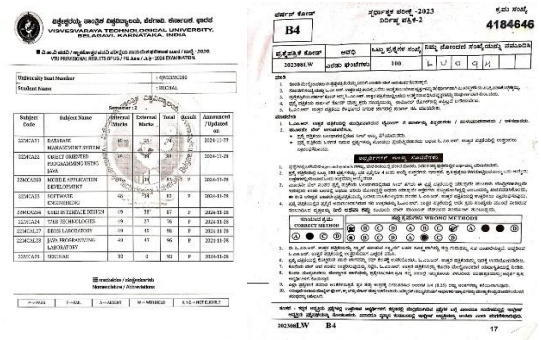


Figure 1. Block Diagram of Proposed system

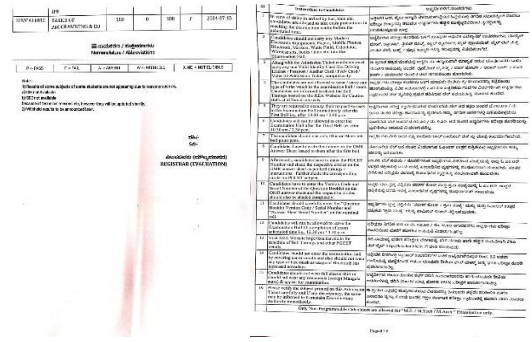
V. RESULT and DISCUSSION

This web application uses Flask and the EasyOCR library to analyze and extract text from images that users upload that contain text in both Kannada and English. As it uploads, the image is processed for

optical character recognition and saved to a precise folder. Using Unicode character ranges, the extracted The text is split up into separate words, which are subsequently classified as either Kannada or English. The application determines the proportion of words in each language and presents the findings on a webpage. The entire extracted text as well as different English and Kannada segments are available for users to view. It helps in determining the linguistic composition of bilingual documents. The application is accessible and easy to use because it allows real-time interaction through the browser.



(a) (b)

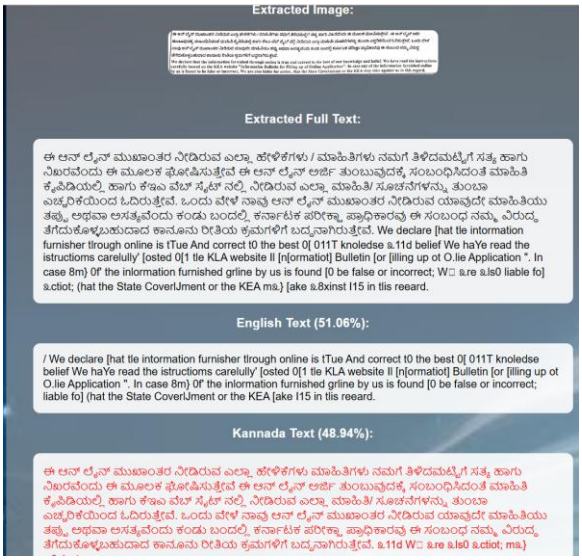


(b) (d)

Figure 2. Bilingual Documents (a), (b), (c), and (d), shows the real-time image of mark cards and government forms as given input.

(a)

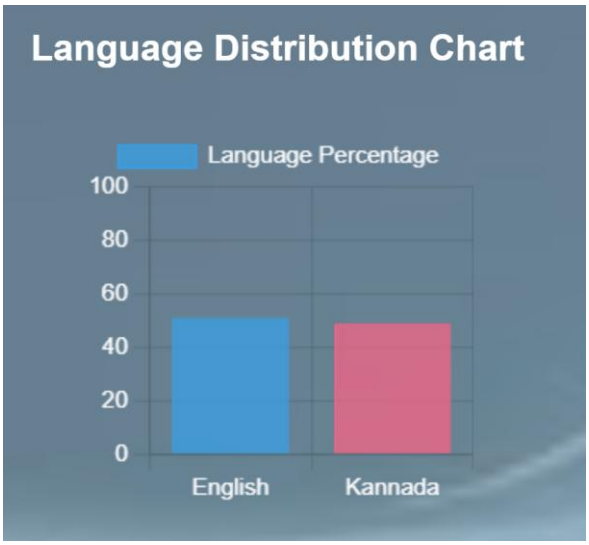
(b)



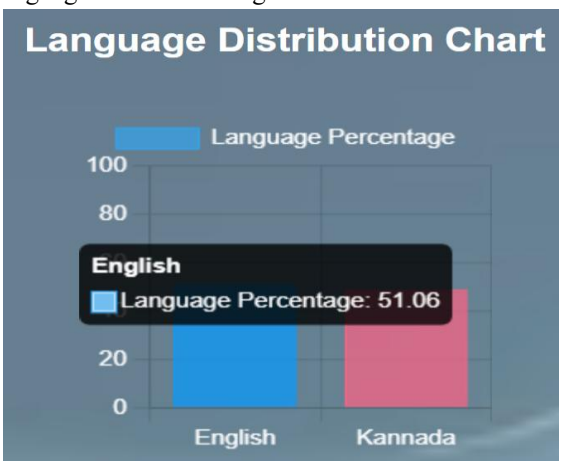
(c)

Figure 3. Text Extraction and Classification (a), (b), and (c) shows the result of bilingual document.

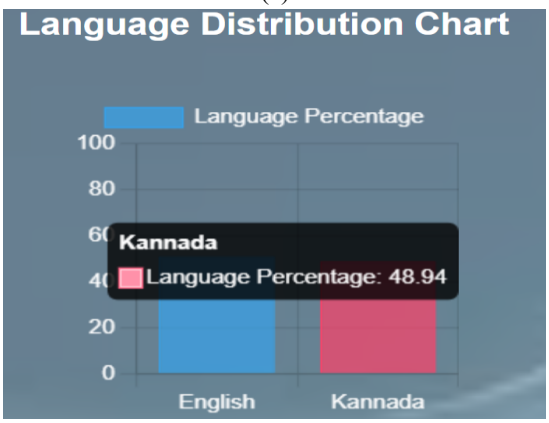
This is how an efficient management of bilingual documents, i.e., forms that belong to government authority or are Karnataka's educational materials, can be made when deployed in both Kannada and English. The technique is free from the complexities of deep NLP by using EasyOCR and Unicode-based classification and still gives an authentic result. The Unicode ranges of both scripts combine, and therefore the basic character-level classification rule can be successfully used on them. Possible extensions would be a preprocessing filter on low-quality images, expanding bilingual inputs, and adding some other form of graphical analysis (a bar chart) to better visualize the proportions of each language.



The language form of a bilingual document is displayed in this bar graph, which contrasts Kannada and English. EasyOCR output for OCR is the basis for the analysis, which is then classified using Unicode ranges. Users are quickly able to understand which language is used in bilingual document.



(a)



(b)

Figure 5. Language Percentage

In Figure 5. (a) and (b) Shows the language distribution in a bilingual document is displayed in these bar charts, with English representing 51.06% and Kannada for 48.94%. The small variation suggests that the extracted text contains a nearly equal amount of both languages.

VI. CONCLUSION

As shown in figure 5, (a) and (b), the language distribution in a bilingual document, the reason behind the 51.06% and 48.94% split rather than a 50%-50% split, has to do with the underlying data for training. For example, if the input data had a bit more content in English relative to Kannada, It would have almost become reflected in the output by the model. Machine learning models make no assumptions about equal representation; rather, they follow the data given. All it takes is even a small difference in the number of English and Kannada samples in the data to produce such results. In essence, the model is using the patterns it has learned to assign probabilities. The English and Kannada categories must add up to 100% since they are non-overlapping. The results show that English content is slightly more prevalent in the evaluated data (51.06% English and 48.94% Kannada), which is entirely normal and expected in a real-world machine learning use case.

The project does extract text accurately from input images which in this case are of Kannada and English using EasyOCR's model CRNN. Simple models of language need to be used in order to avoid more sophisticated natural language models. Its modular nature of the system incorporating image upload, OCR, categorization and visualization. EasyOCR proposes a better overall result for identifying mixed-language (English and Kannada) text in images from a practical perspective. It employs a simple Unicode-based technique to differentiate between languages, is more accurate, and manages font and layout variations better. While Tesseract is still functional, it is generally less dependable for Kannada and frequently requires additional preprocessing or cleanup to produce results that are functioning.

Feature	Easy OCR (Approx % Accuray)	Tesseract OCR (Approx % Accuracy)	Remarks
English	95%	85%	Tesseract performs well in English,

			while EasyOCR is more consistent across fonts and layouts.
Kannada	90%	60%	EasyOCR detects Kannada letters and words more accurately.
Language Classification Accuracy	98%	75%	EasyOCR implements accurate Unicode recognition, whereas Tesseract's langdetect can misclassify
Bilingual Handling	95%	65%	EasyOCR differentiates English and Kannada more accurate.

Figure 6. Contrast Table

In Figure 6, this has been concluded that EasyOCR is preferable when working with bilingual documents, particularly in regional scripts like Kannada.

REFERENCE

[1] Dr. S. Basavaraj Patil, *"Neural Network based Bilingual OCR System: Experiment with English and Kannada Bilingual Documents.",* International journal of Computer Applications, vol. 13-No8, 2011.

[2] Santhosh S. G. and Sridhara Acharya P, *"A Survey on Bilingual answer script extraction from low- resolution images."*, International journal of progressive research in engineering management and science (IJPREMS), vol. 04, Issue 08, pp. 416-421, 2024.

[3] Shashank Simha B. K, Rahul M, Jyoti R. Munavalli, Prajwal Anand, *"Dual Language Detection using Machine Learning."*, International Conference on VLSI, Communications and Computer Communication, Advances in Intelligent Systems and Technologies, DOI: 10.53759/aist/978-9914-9946-1-2_32, 2023.

[4] M. Liao, Z. Zou, Z. Wan, C. Yao, and X. Bai, *"Real-time scene text detection with differentiable binarization and adaptive scale fusion,"*. IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 45, no. 1, pp. 919–931, 2023. DOI: 10.1109/TPAMI.2022.3155612

[5] Sammed S. Admuthe, Hemlata P. Channe, *"Document Image Classification using Visual and Textual Features"*. International Journal of Engineering Research & Technology (IJERT) Vol. 10 Issue 09, September-2021.

[6] Fazeeia Mohammed, Patrick Hosein, Aqeel Mohammed, C. Kiran Mai, *"Bilingual Dialect Classification using NLP",* International Conference on Advances in Computer Engineering and Communication Systems (ICACECS 2023), Atlantis Highlights in Computer Sciences 18, DOI 10.2991/978-94-6463-314-6_5, 2023.

[7] Shivani Surana, Komal Pathak, Vidhan Shrivastava, Mahesh T R. *"Text Extraction and Detection from Images using Machine Learning Techniques: A Research Review"*, Proceedings of the International Conference on Electronics and Renewable Systems (ICEARS 2022), DOI: 10.1109/ICEARS53579.2022.9752274, 2022.

[8] Sanjana Gunna, Rohit Saluja, C. V. Jawahar. *"Transfer learning for Scene Text Recognition in Indian Languages"*, International Conference on Document Analysis and Recognition (ICDAR) 2021: Document Analysis and Recognition, ICDAR 2021 Workshops, pp 182-197, 2022.

[9] Shaswata Saha, Neelotpal Chakraborty, Soumyadeep Kundu, Sayantan Paul, Ayatullah Faruk Mollah, Subhadip Basu, Ram Sarkar, *"Multi-lingual scene text detection and language identification",* International Association for Pattern Recognition, Volume 138, Pages 16-22, 2020.

[10] Shalini Puri, Satya Prakash Singh, *"Advanced Applications on Bilingual Document Analysis and Processing Systems",* International Journal of Applied Metaheuristic Computing, Volume 11, Issue 4, 2020.

[11] C. K. Ch'ng and C. S. Chan, "*Total-text: A comprehensive dataset for scene text detection and recognition*", Proc. Int. Conf. Document Anal. Recognition, pp. 935-942, 2017.

[12] V.N.M. Aradhya, and M.S. Pavithra, *"A Comprehensive of Transforms, Gabor filter and k-means Clustering for Text Detection in Images*

and Video," Applied Computing and Informatics Vol. 12, pp. 109–116, 2016.

[13] S. Yousfi, A. Berrani, and C. Garcia, "*Boosting based Approaches for Arabic Text Detection in News Videos,*" in 11th IAPR International Workshop on Document Analysis Systems (DAS'14), Tours, France,2014.

[14] Le Kang, Jayant Kumar, Peng Ye, Yi Li, and David Doermann, "*Convolutional Neural Networks for Document Image Classification*", in International Conference on Pattern Recognition (ICPR), 2014.