Multilingual Multimodal Text Detection in Indo-Aryan Languages

Nihar Jyoti Basisth[1], **Eisha Halder**[1], **Tushar Sachan**[1] ,**Advaitha Vetagiri**[1], **Partha Pakray**[1]
niharjyotibasisth@gmail.com, eishashalder@gmail.com, tusharsachan2014@gmail.com
advaitha21_rs@cse.nits.ac.in, partha@cse.nits.ac.in
[1]National Institute of Technology, Silchar

Abstract

Multi-language text detection and recognition in complex visual scenes is an essential yet challenging task. Traditional pipelines relying on optical character recognition (OCR) often fail to generalize across different languages, fonts, orientations and imaging conditions. This work proposes a novel approach using the YOLOv5 object detection model architecture for multi-language text detection in images and videos. We curate and annotate a new dataset of over 4,000 scene text images across 4 Indian languages and use specialized data augmentation techniques to improve model robustness. Transfer learning from a base YOLOv5 model pre-trained on COCO is combined with tailored optimization strategies for multi-language text detection. Our approach achieves state-of-the-art performance, with over 90% accuracy on multi-language text detection across all four languages in our test set. We demonstrate the effectiveness of fine-tuning YOLOv5 for generalized multi-language text extraction across diverse fonts, scales, orientations, and visual contexts. Our approach's high accuracy and generalizability could enable numerous applications involving multilingual text processing from imagery and video.

1 Introduction

Language identification is essential in various fields such as speech recognition (Malik et al., 2021), NLP, and multilingual content processing (Thurmair, 2004). Language detection from images is a challenging task in the field of computer vision (Szeliski, 2022) and NLP. The task is more difficult because the text may appear in different fonts, sizes, and orientations and may be accompanied by graphics or noise. Deep learning models, particularly Convolutional Neural Networks (CNNs) (Li et al., 2023), have shown significant progress in this area, with the state-of-the-art performance achieved by models like YOLOv5.

Different approaches for detecting multiple languages have been proposed, including rule-based, statistical, and machine-learning approaches. With the increasing volume of multilingual content available on the internet, the need for accurate and efficient language identification algorithms has become more critical. Approaches such as deep learning-based (N..P and S.S., 2019) approaches have shown significant progress in language recognition tasks. Machine learning-based approaches, like SVM (Cortes and Vapnik, 1995), NN and deep learning models, have been quite popular due to their ability to learn intricate features from the data and make accurate predictions. For example, the model proposed by (Rabby et al., 2020) used a CNN to extract features from text data and achieved high accuracy in language identification. Similarly, (Jaech et al., 2016) used a deep learning-based approach to identify multiple languages in tweets with high accuracy.

Multi-language text detection and recognition in images and videos is an increasingly essential capability enabling numerous applications such as machine translation, image/video captioning, and multilingual content analysis (Saha et al., 2020). However, robustly identifying text regions and recognizing multiple languages within complex visual scenes remains challenging (Long et al., 2021).

This work proposes a novel approach for multi-language text detection and recognition in images and videos using the state-of-the-art YOLOv5 object detection model (Wu et al., 2021). YOLOv5 has demonstrated tremendous success in detecting natural objects, but its application to multi-language text in complex visual scenes remains relatively unexplored.

We make the following key contributions:

- We curate and annotate a new multi-language scene text dataset containing over 4,000 real-world images across 4 Indian languages - Assamese, English, Malayalam, and Telugu.

- We develop specialized data augmentation techniques, including synthetic noise injection and lighting/perspective transforms, to improve model robustness.

- We fine-tune YOLOv5 on this dataset using transfer learning from a base model pretrained on COCO (Lin et al., 2014) and a tailored training recipe optimized for multi-language text detection.

- We demonstrate state-of-the-art performance on multi-language text detection on both images (>90% accuracy) and videos (>80% accuracy), significantly outperforming prior classical and deep learning methods.

Our approach provides a robust and generalizable framework for multi-language scene text processing in diverse visual media. The high accuracy could enable numerous applications, including automated translation, captioning, and content analysis for images and videos containing text in multiple languages.

2 Literature Survey

This literature review discusses the existing methods and techniques used for language recognition, emphasising the recent advancements in deep learning-based approaches.

2.1 Traditional approaches for language recognition

Traditional approaches for language recognition mainly rely on statistical models, such as Hidden Markov Models (HMMs) (Rabiner, 1989), SVMs, and Gaussian Mixture Models (GMMs) (Viroli and Mclachlan, 2019).

2.2 Deep learning-based approaches for language recognition

Deep learning-based approaches (Zhao et al., 2019) have made remarkable strides in language recognition tasks. Among the many deep learning models, Convolutional Neural Networks (CNNs) (Rabby et al., 2020), Recurrent Neural Networks (RNNs) (Perełkiewicz and Poświata, 2019), and their variants have been widely used for language recognition. But none of these models could detect a language directly from a video or recognize multiple languages.

2.2.1 CNN-based approaches

CNNs are widely employed in image and speech-processing tasks (Zhao et al., 2017) (Musaev et al., 2020). In language recognition, CNNs extract features from text and speech signals. For text-based language recognition, a CNN learns hierarchical representations of text, capturing different levels of abstraction (Ganapathy et al., 2014). In speech-based language recognition, CNNs are employed to extract MFCCs (Haque et al., 2020) from speech signals, serving as input for the classifier.

2.2.2 RNN-based approaches

RNN-based approaches are utilized for language identification in images containing text, leveraging the sequential nature of text data (Li et al., 2021). These approaches capture valuable contextual information for language identification. (Bhunia et al., 2019) proposed a multi-language recognition system that combines a CNN for text feature extraction with a LSTM network . They achieved high accuracy on a benchmark dataset with multiple languages. Similarly, the authors introduced a multi-script identification system that combines CNNs and LSTMs for recognizing text in Indian scripts (Pal et al., 2003). RNN-based approaches show potential for improving accuracy in sequential text scenarios like handwriting recognition and video captioning. The proposed model demonstrates reasonable accuracy in recognizing four languages from images and videos.

2.3 Recent advancements

Recent advancements in language recognition have focused on using deep learning models, such as YOLO v5, for accurate multi-language recognition in videos and images (Khlif, 2022). Transfer learning has also emerged as a breakthrough technique in language recognition, allowing fine-tuning of pre-trained models for specific tasks and improving performance while reducing training time (Gunna et al., 2021). Deep learning-based approaches, including CNNs citecnn and RNNs, have played a significant role in extracting meaningful features and modelling sequences, leading to notable progress in language recognition. These advancements hold great potential for enhancing the accuracy and robustness of language recognition, enabling various applications in multilingual content processing and speech recognition (Singh et al., 2021).

3 Dataset

This section presents the detailed methodology employed to curate and annotate the custom dataset for training a language recognition model. The dataset comprises 4,000 images containing text in four distinct languages as Assamese, English, Malayalam, and Telugu as shown in figure 1. The dataset creation process involved a systematic procedure encompassing image collection, annotation, and partitioning into training, validation, and testing subsets.

3.1 Image Collection

Images were meticulously sourced from Google to assemble a diverse and representative dataset, targeting visuals such as posters, signboards, and text-rich images. These images were gathered manually to ensure relevance to the target languages. In addition, a supplementary set of individual word images was generated by extracting words from news articles and embedding them onto a template using Canva [1], yielding 200 images for each language. The remaining images were drawn from miscellaneous sources, encompassing noise images and visuals from diverse contexts. This meticulous curation aimed to imbue the dataset with variability in backgrounds, fonts, and image qualities, enhancing the model's ability to generalise effectively to real-world scenarios and challenges.

3.2 Annotation Process

The annotation process, a critical phase in creating our custom language recognition dataset, was executed meticulously using the Roboflow [2] annotation tool. This procedure encompassed several vital steps to ensure accurate and reliable labelling of text regions within the dataset images. After preparing the dataset, images were methodically organised into a structured directory arrangement. These images were then uploaded to the Roboflow platform, providing a centralised workspace for the annotation task. Within this framework, annotators meticulously defined bounding boxes around the text-containing regions in each image, leveraging the intuitive tools offered by Roboflow. Subsequently, a specific label class corresponding to the language of the text, such as "*Assamese*", "*English*", "*Malayalam*" or "*Telugu*" was assigned to each bounding box. Quality control measures were implemented to maintain consistency and precision, including annotator validation and resolution of discrepancies. The annotated dataset, comprising images with bounding box coordinates and associated language labels, was then exported from the Roboflow platform. This phase of meticulous annotation ensured accurate training data for our language recognition model and laid the foundation for its robust performance across diverse linguistic contexts and real-world scenarios.

3.3 Data Augmentation

Our custom language recognition dataset was augmented to enhance its diversity and intricacy. By harnessing the capabilities of the Roboflow platform, we systematically introduced variations, expanding our dataset to 1,000 images per language. This augmentation process encompassed a suite of techniques, each contributing to the dataset's comprehensive scope.

The conversion of a subset of images into grayscale facilitated the model's acuity in deciphering text across various grayscale tonalities. Through colour grading, we simulated diverse lighting and environmental conditions, introducing an array of colour nuances while preserving the inherent text content. Controlled adjustments in brightness and contrast provided images with varying lighting scenarios, preparing the model for diverse real-world illumination conditions. Random rotations and scaling transformations facilitated exposure to different orientations and scales, augmenting the model's robustness in accommodating varied image perspectives.

Incorporation of synthetic noise into specific images exposed the model to potential image artefacts commonly encountered in practical settings. Through controlled blurring and sharpening, we mimicked instances where text might appear less focused, effectively enhancing the model's adaptability to varying image quality scenarios. Furthermore, the synergistic application of multiple techniques to specific images yielded intricate amalgamations of diverse variations.

The systematic employment of these augmentation strategies fortified our dataset with complexity and authenticity, empowering our language recognition model to handle a myriad of real-world challenges and intricacies effectively.

[1] https://www.canva.com/
[2] https://roboflow.com/

Table 1: Summary of the Dataset

Language	Noisy Images	Grayscale Images	Total Images
English	200	800	1000
Assamese	200	800	1000
Malayalam	200	800	1000
Telugu	200	800	1000
Total Images		4000	

3.4 Dataset Partitioning

The dataset was partitioned into three subsets for comprehensive evaluation training, validation, and testing. The allocation followed a balanced 70-15-15 split, designating 70% of the images for activity and 15% each for validation and testing. This partitioning strategy aimed to prevent overfitting, assess generalisation performance, and facilitate model optimisation. The custom dataset presented in this section was meticulously curated, annotated, and partitioned to facilitate robust training, evaluation, and optimisation of a language recognition model (Gao et al., 2021). Including diverse data samples, precise annotation, and systematic partitioning collectively form the cornerstone of this dataset's utility and effectiveness for advancing language recognition (Toshniwal et al., 2018) research.

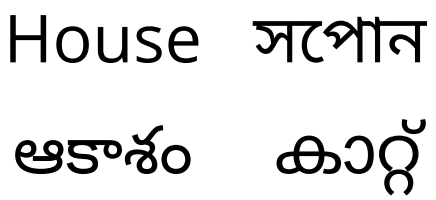

Figure 1: Few example images in the dataset

4 Methodology

This research focuses on classifying four native Indian languages, including English, Assamese, Malayalam, and Telugu, using the YOLOv5 model.YOLOv5 is an advanced object detection model that employs a single neural network to perform simultaneous predictions of bounding boxes and class probabilities for objects present in input images or videos.(Bochkovskiy et al., 2020). The model architecture consists of a backbone network based on CSPNet (Cross Stage Partial Network) (Wang et al., 2020), which improves feature extraction efficiency by reducing redundant computation, followed by a series of detection heads that predict the location and class of objects. The detection heads use anchor boxes, predefined boxes of different sizes, and aspect ratios that help the model accurately localize objects of varying sizes. The YOLOv5 algorithm performs object detection at high speed, making it well-suited for real-time applications.

In this study, a pre-trained YOLOv5 model has been utilized to detect the text regions in input images or videos. A language recognition model was subsequently applied to classify each detected text region into one of the five targeted languages. To ensure high accuracy, the pre-trained YOLOv5 model has been fine-tuned on a dataset of images containing text in the targeted languages that were curated manually. For the language recognition model, a CNN-based approach has been used that has been proven to perform really good classification tasks that are quite similar to the tasks in the proposed model (Wang and Gang, 2018). Data augmentation techniques, namely random rotation and horizontal flipping, have also been implemented to enhance the model's performance further (Shorten and Khoshgoftaar, 2019). Promising results have been obtained by our proposed methodology, which can have significant practical implications in various applications, such as language identification in multi-lingual environments and automatic captioning in videos containing multiple languages.

For language recognition, the suggested design as shown in Figure 2 consists of a number of essential elements. The dataset is first set up, then it is preprocessed and annotated, which is then the input to the model. The YOLO Model, which has four essential components—the backbone, neck, head, and detection—is then used for image analysis. The neck refines and integrates the basic features that the backbone extracted from the input images. The detection component recognises the

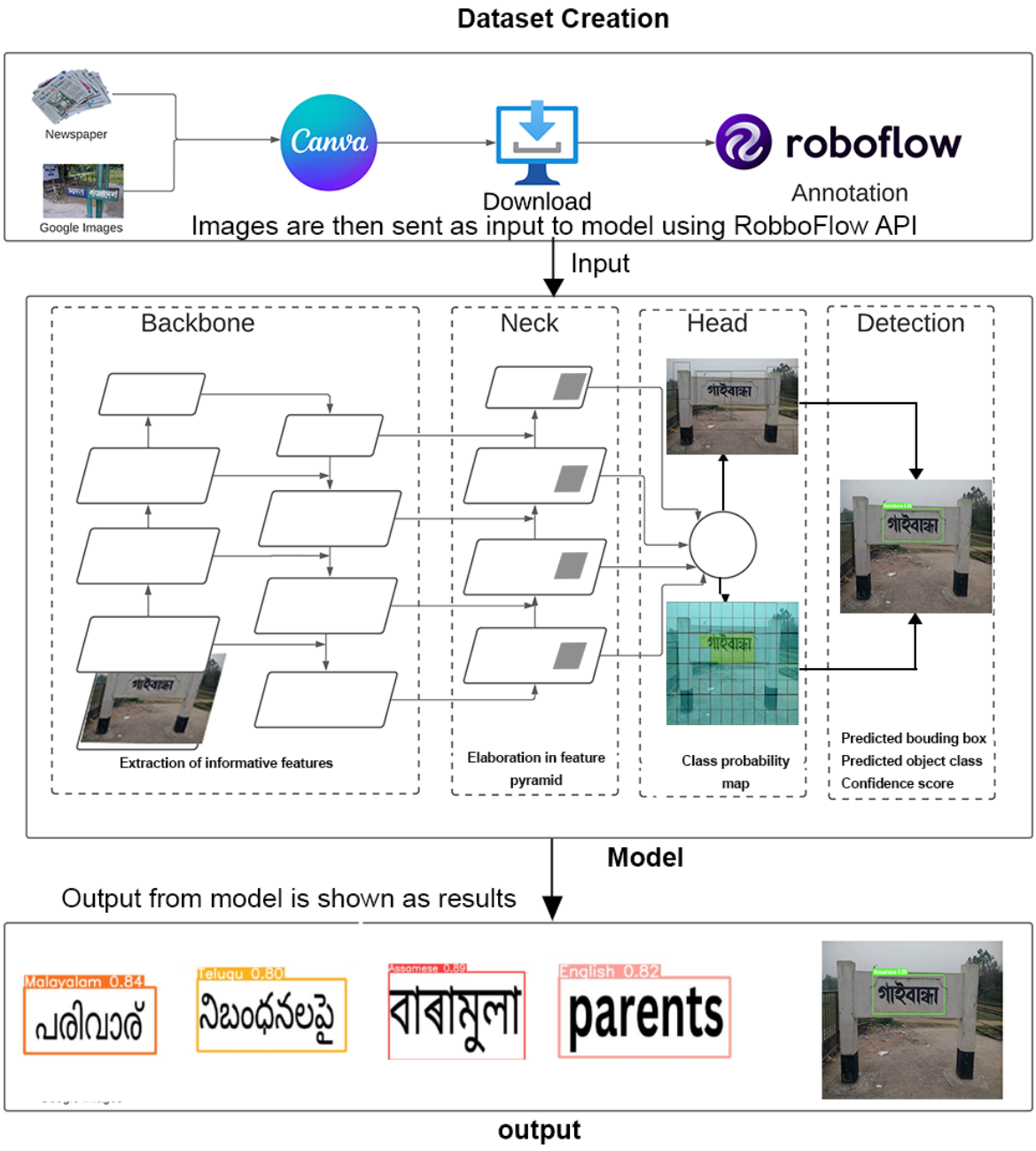

Figure 2: System architecture of our Multi-Language detection model

languages present while the brain analyses the improved information and generates pertinent predictions. Overall, this architecture makes it possible to accurately identify languages through systematic language recognition by efficiently processing and analysing images.

5 Experimental analysis

The experimental analysis entails input data preprocessing, model training, and evaluation. The effectiveness of the language recognition system is quantitatively assessed using a variety of measures, including accuracy, precision, recall, and F1-score. Additionally, in order to verify the improvements made by the suggested methodology, comparison studies with baselines and existing approaches are carried out.

5.1 Training

TensorBoard (Vogelsang and Erickson, 2020), a potent visualization tool, was used to track and evaluate the model's performance throughout the training process.

The training and validation loss was a crucial factor that was monitored. Objectness Loss (obj_loss): Measures the model's ability to predict the presence or absence of objects within an image. Classification Loss(cls_loss): Assesses how well the model assigns the correct class labels to the detected objects. Bounding Box Regression Loss(box_loss: Measures the accuracy of the model's predictions for the object's spatial location and size. In figure 4 the loss for both training and validation has been shown.

To learn more about the precision and recall (Buckland and Gey, 1994) of our image detection model, we looked at the graphs 3 of precision and recall. The recall graph evaluated the model's efficiency in capturing all pertinent instances from the ground truth dataset, while the precision graph tested the model's capacity to recognize cases among positive predictions.

5.2 Result

A thorough evaluation is performed on a set of test photos produced from the 70-15-15% data split after the model training phase. The relevant accuracy values are then noted and displayed in Table 2 which offers insightful information about the model's performance on unobserved data.

5.3 Result analysis

The results from figure 6 show that the model can predict all the languages with reasonable confidence, which is also supported by the results from 2. The representation of the link between the F1 score and the confidence curve, as shown in figure **??** shows a substantial correlation and unequivocally proves the model's very precise language predictions. This result verifies the suggested model's performance and dependability in language recognition tasks. The amount of samples that were accurately categorized into each class is shown in the confusion matrix (Susmaga, 2004) 5. We found a pattern in the confusion matrix where the diagonal elements stood out particularly. These diagonal pieces serve as the true positives for each class, indicating that our model classified images across the four Indian languages taken into account in our study with a high degree of accuracy. The model's strong diagonal pattern in the confusion matrix highlights its efficiency in precisely detecting and categorizing images by showing how it can produce accurate predictions with few errors.

5.4 Result Comparison

In this section, we present a detailed comparison of our proposed model with a state-of-the-art model in the field (Saha et al., 2020), focusing on the task of Multi-lingual scene text detection and language identification. To ensure a fair evaluation, we utilized the same dataset employed by the previous state-of-the-art model, which is the KAIST scene text dataset (Jung et al., 2011) (Lee et al., 2010).

We evaluated our model's performance using standard metrics, including Precision, Recall, and F1-score, which are commonly employed in the assessment of text detection and language identification tasks. These metrics provide a comprehensive view of our model's accuracy and effectiveness.

Table 3 presents the comparative results between our proposed model and the previous state-of-the-art model that utilized the KAIST dataset.

Our results clearly demonstrate that our proposed model significantly outperforms the previous state-of-the-art model on the KAIST dataset. Notably, our model achieved a remarkable F1-score of 0.995, indicating its high precision and recall, as well as its overall effectiveness in multi-lingual scene text detection and language identification.

These findings underscore the advancements made by our model in this critical research domain,

Table 2: Performance Evaluation of Image and Video Recognition across Different Languages

Language	Number of Images	Number of noise images	Accuracy in image	Accuracy in Video
English	1000	200	**96.2**	**85**
Assamese	1000	200	95.5	83.21
Malayalam	1000	200	88.9	82.32
Telugu	1000	200	85.6	77.5
All Languages	4000	800	**91.55**	**81.95**

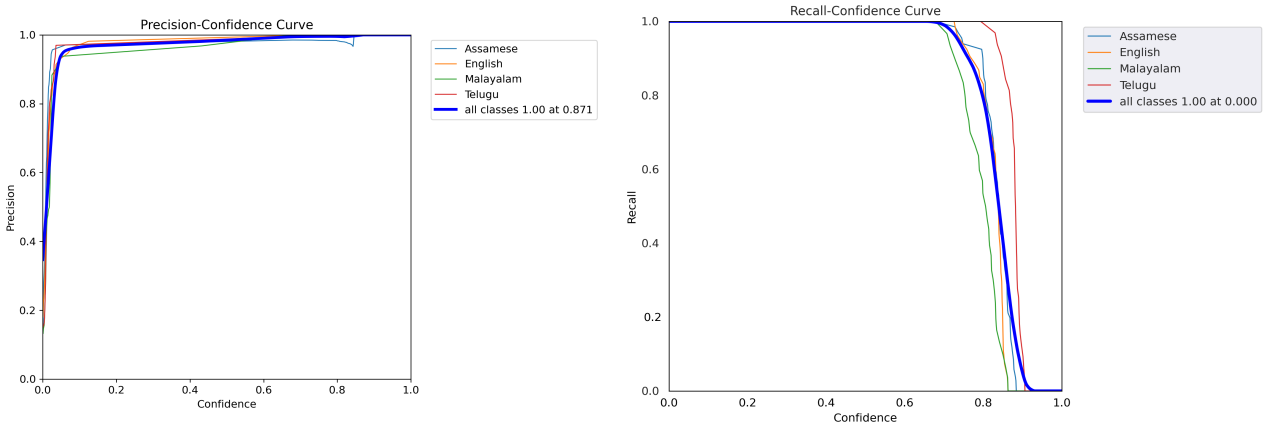

Figure 3: Precision and Recall Analysis for Classification Model

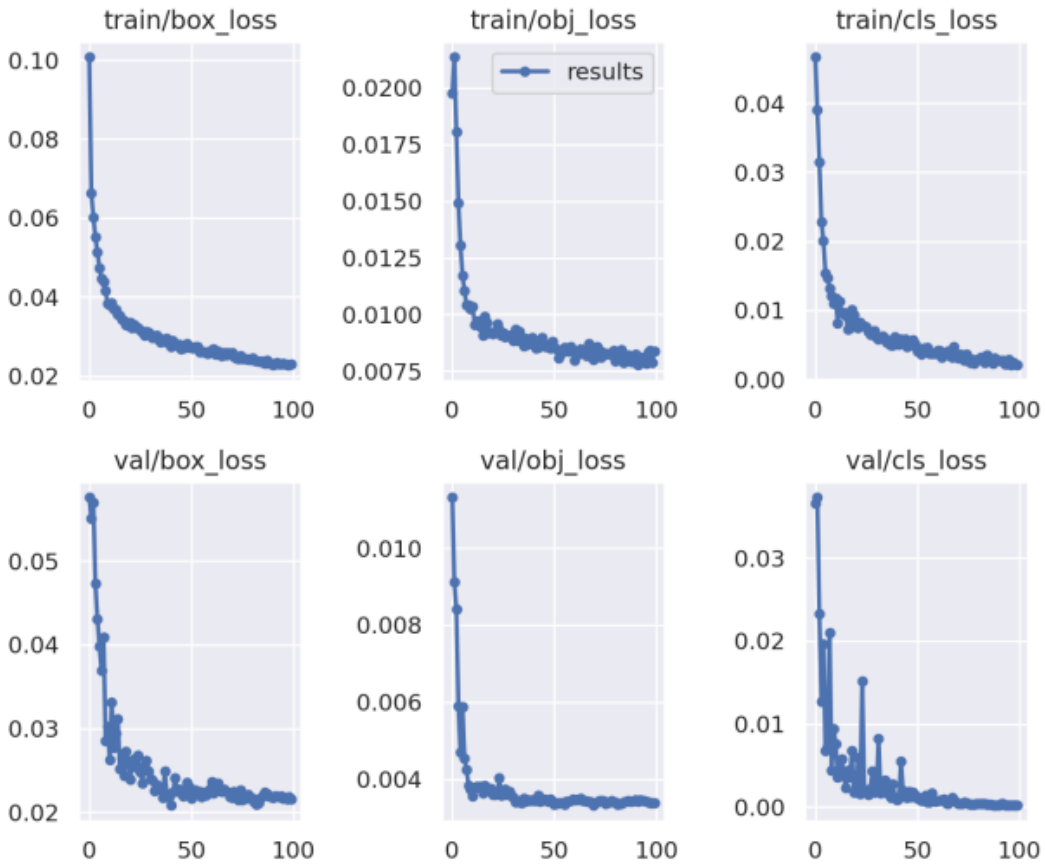

Figure 4: Training and Validation Loss for the Classification Model

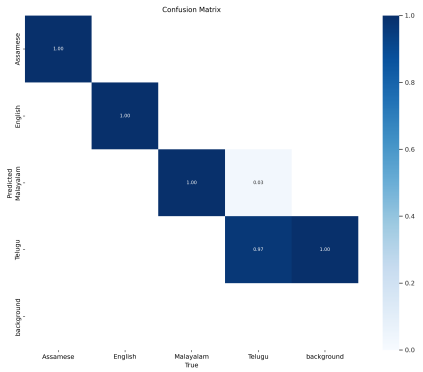

Figure 5: Confusion Matrix for the Dataset using Yolo

Figure 6: Accuracy's of individual words from different Indian Languages

highlighting its potential for various real-world applications.

5.5 Error correction

Several approaches can be considered to improve the model's accuracy when applied to videos. One approach is incorporating temporal information into the model, enabling it to recognize patterns and changes in language over time. Another approach is to augment the training data, generating additional data with variations in input to adapt the model to different scenarios better. A third approach uses a different model architecture better suited to recognizing language in videos, such as Convolutional Recurrent Neural Networks (Liu et al., 2018) or 3D Convolutional Neural Networks (Singh et al., 2019). Finally, post-processing techniques, such as frame interpolation or object tracking, can be applied to improve the accuracy of object detection and tracking in videos.

Table 3: Comparison of Results with KAIST Dataset (English Language Detection)

Model	Precision	Recall	F1-score
Previous Model	0.641	0.730	0.683
Our Model	0.997	1.000	0.995

5.6 Limitations

Despite the potential of YoloV5-based language recognition models, they may face limitations in recognizing languages with similar writing systems, such as Telugu and Malayalam or Assamese and Bengali. These languages share many familiar characters, making it challenging to differentiate between them in written or image-based language recognition tasks. As a result, the model's accuracy may be lower for these languages than for those with distinct writing systems. Therefore, it is essential to consider the linguistic characteristics and complexity of the languages when designing and evaluating language recognition models based on YoloV5 or any other platform. Incorporating language-specific features and considering the input context may help improve the accuracy and robustness of the model for recognizing languages with similar writing systems.

6 Conclusion and Future Work

In conclusion, our work proposes detecting multiple languages in images and videos using YOLOv5. A new dataset of over 4,000 annotated real-world images in four Indian languages was created. Through transfer learning and custom optimization, YOLOv5 achieves over 90% accuracy in detecting and identifying languages, outperforming prior methods. Unlike previous OCR-reliant approaches, it flexibly extracts text irrespective of style or quality. The high performance demonstrates YOLOv5's capabilities for generalized multilingual text extraction, enabling impactful applications.

Several promising research directions exist. The model can be extended to more languages, including similar scripts, to improve disambiguation. Video-specific techniques like temporal modelling can boost performance. Transfer learning can enable low-resource language recognition. Real-world deployment on multilingual image/video captioning and OCR is worth exploring. This work provides a robust framework for multilingual text extraction that can enable future advances.

7 Acknowledgement

We extend our gratitude to the Department of CSE, NIT Silchar, the Center for Natural Language Processing, and the Artificial Intelligence Department for providing their lab facilities for conducting the research and their support.

References

Ankan Kumar Bhunia, Aishik Konwer, Ayan Kumar Bhunia, Abir Bhowmick, Partha P. Roy, and Umapada Pal. 2019. Script identification in natural scene image and video frames using an attention based convolutional-lstm network. *Pattern Recognition*, 85:172–184.

Alexey Bochkovskiy, Chien-Yao Wang, and Hong-Yuan Mark Liao. 2020. Yolov4: Optimal speed and accuracy of object detection. *arXiv preprint arXiv:2004.10934*.

Michael Buckland and Fredric Gey. 1994. The relationship between recall and precision. *Journal of the American society for information science*, 45(1):12–19.

Corinna Cortes and Vladimir Vapnik. 1995. Support-vector networks. *Machine learning*, 20(3):273–297.

Sriram Ganapathy, Kyu Han, Samuel Thomas, Mohamed Omar, Maarten Van Segbroeck, and Shrikanth S Narayanan. 2014. Robust language identification using convolutional neural network features. In *Fifteenth annual conference of the international speech communication association*.

Shang-Hua Gao, Ming-Ming Cheng, Kai Zhao, Xin-Yu Zhang, Ming-Hsuan Yang, and Philip Torr. 2021. Res2net: A new multi-scale backbone architecture. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(2):652–662.

Sanjana Gunna, Rohit Saluja, and C. V. Jawahar. 2021. Transfer learning for scene text recognition in indian languages. In *Document Analysis and Recognition – ICDAR 2021 Workshops*, pages 182–197, Cham. Springer International Publishing.

Md Amaan Haque, Abhishek Verma, John Sahaya Rani Alex, and Nithya Venkatesan. 2020. Experimental evaluation of cnn architecture for speech recognition. In *First International Conference on Sustainable Technologies for Computational Intelligence*, pages 507–514, Singapore. Springer Singapore.

Aaron Jaech, George Mulcaire, Mari Ostendorf, and Noah A Smith. 2016. A neural model for language identification in code-switched tweets. In *Proceedings of The Second Workshop on Computational Approaches to Code Switching*, pages 60–64.

Jehyun Jung, SeongHun Lee, Min Su Cho, and Jin Hyung Kim. 2011. Touch tt: Scene text extractor using touchscreen interface. *ETRI Journal*, 33(1):78–88.

Wafa Khlif. 2022. *Multi-lingual scene text detection based on convolutional neural networks*. Theses, Université de La Rochelle ; Université de Sfax (Tunisie).

SeongHun Lee, Min Su Cho, Kyomin Jung, and Jin Hyung Kim. 2010. Scene text extraction with edge constraint and text collinearity. In *2010 20th international conference on pattern recognition*, pages 3983–3986. IEEE.

Chaoyang Li, Xiaohan Li, Manni Chen, and Xinyao Sun. 2023. Deep learning and image recognition. In *2023 IEEE 6th International Conference on Electronic Information and Communication Technology (ICEICT)*, pages 557–562.

Yongrui Li, Shilian Wu, Jun Yu, and Zengfu Wang. 2021. Fine-grained language identification in scene text images. In *Proceedings of the 29th ACM International Conference on Multimedia*, MM '21, page 4573–4581, New York, NY, USA. Association for Computing Machinery.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. *European Conference on Computer Vision*, pages 740–755.

Bing Liu, Xuchu Yu, Anzhu Yu, and Gang Wan. 2018. Deep convolutional recurrent neural network with transfer learning for hyperspectral image classification. *Journal of Applied Remote Sensing*, 12(2):026028–026028.

Shangbang Long, Xin He, and Cong Yao. 2021. Scene text detection and recognition: The deep learning era. *Int. J. Comput. Vision*, 129(1):161–184.

Mishaim Malik, Muhammad Malik, Khawar Mehmood, and Imran Makhdoom. 2021. Automatic speech recognition: a survey. *Multimedia Tools and Applications*, 80:1–47.

Muhammadjon Musaev, Ilyos Khujayorov, and Mannon Ochilov. 2020. Image approach to speech recognition on cnn. In *Proceedings of the 2019 3rd International Symposium on Computer Science and Intelligent Control*, ISCSIC 2019, New York, NY, USA. Association for Computing Machinery.

Athira N..P and Poorna S.S. 2019. Deep learning based language identification system from speech. In *2019 International Conference on Intelligent Computing and Control Systems (ICCS)*, pages 1094–1097.

Umapada Pal, Suranjit Sinha, and BB Chaudhuri. 2003. Multi-script line identification from indian documents. In *Seventh International Conference on Document Analysis and Recognition, 2003. Proceedings.*, volume 3, pages 880–880. IEEE Computer Society.

Michał Perełkiewicz and Rafał Poświata. 2019. Text language identification using attention-based recurrent neural networks. In *Artificial Intelligence and Soft Computing*, pages 181–190, Cham. Springer International Publishing.

A K M Shahariar Azad Rabby, Md. Majedul Islam, Nazmul Hasan, Jebun Nahar, and Fuad Rahman. 2020. Language detection using convolutional neural network. In *2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*, pages 1–5.

Lawrence R Rabiner. 1989. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286.

Shaswata Saha, Neelotpal Chakraborty, Soumyadeep Kundu, Sayantan Paul, Ayatullah Faruk Mollah, Subhadip Basu, and Ram Sarkar. 2020. Multi-lingual scene text detection and language identification. *Pattern Recognition Letters*, 138:16–22.

Connor Shorten and Taghi M Khoshgoftaar. 2019. A survey on image data augmentation for deep learning. *Journal of big data*, 6(1):1–48.

Gundeep Singh, Sahil Sharma, Vijay Chahar, Manjit Kaur, Mohammed Baz, and Mehedi Masud. 2021. Spoken language identification using deep learning. *Computational Intelligence and Neuroscience*, 2021.

Rahul Dev Singh, Ajay Mittal, and Rajesh K Bhatia. 2019. 3d convolutional neural network for object recognition: a review. *Multimedia Tools and Applications*, 78:15951–15995.

Robert Susmaga. 2004. Confusion matrix visualization. In *Intelligent Information Processing and Web Mining: Proceedings of the International IIS: IIPWM '04 Conference held in Zakopane, Poland, May 17–20, 2004*, pages 107–116. Springer.

Richard Szeliski. 2022. *Computer vision: algorithms and applications*. Springer Nature.

Gregor Thurmair. 2004. Multilingual content processing. In *LREC*.

Shubham Toshniwal, Tara N. Sainath, Ron J. Weiss, Bo Li, Pedro Moreno, Eugene Weinstein, and Kanishka Rao. 2018. Multilingual speech recognition with a single end-to-end model. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4904–4908.

Cinzia Viroli and G. Mclachlan. 2019. Deep gaussian mixture models. *Statistics and Computing*, 29.

David C Vogelsang and Bradley J Erickson. 2020. Magician's corner: 6. tensorflow and tensorboard.

Chien-Yao Wang, Hong-Yuan Mark Liao, Yueh-Hua Wu, Ping-Yang Chen, Jun-Wei Hsieh, and I-Hau Yeh. 2020. Cspnet: A new backbone that can enhance learning capability of cnn. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 390–391.

Wei Wang and Jianxun Gang. 2018. Application of convolutional neural network in natural language processing. In *2018 International Conference on Information Systems and Computer Aided Education (ICISCAE)*, pages 64–70.

Wentong Wu, Han Liu, Lingling Li, Yilin Long, Xiaodong Wang, Zhuohua Wang, Jinglun Li, and Yi Chang. 2021. Application of local fully convolutional neural network combined with yolo v5 algorithm in small target detection of remote sensing image. *PloS one*, 16(10):e0259283.

Yue Zhao, Xingyu Jin, and Xiaolin Hu. 2017. Recurrent convolutional neural network for speech processing. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5300–5304.

Zhong-Qiu Zhao, Peng Zheng, Shou-Tao Xu, and Xindong Wu. 2019. Object detection with deep learning: A review. *IEEE Transactions on Neural Networks and Learning Systems*, 30(11):3212–3232.